

```

do i1=1,4
  j(1)=i1
  do i2=1,4
    j(2)=i2
    do i3=1,4
      j(3)=i3
      do i4=1,4
        j(4)=i4
        if (j(1) .eq. j(2) .or. j(1) .eq. j(3) .or. j(1) .eq. j(4)) cycle
        if (j(2) .eq. j(3) .or. j(2) .eq. j(4)) cycle
        if (j(3) .eq. j(4)) cycle
        print*,j(1),j(2),j(3),j(4)
      end do
    end do
  end do
end do

```

# Journal of Modern Applied Statistical Methods

## *Invited Articles*

202 - 227	<b>Ronald C. Serlin</b>	Constructive Criticism
228 - 239	<b>Ralph D'Agostino, Sr., Lisa M. Sullivan</b>	Chronic Disease Data And Analysis: Current State Of The Field
240 - 242	<b>Thomas R. Knapp</b>	Some Reflections On Significance Testing
243 - 247	<b>Phillip I. Good</b>	Extensions Of The Concept Of Exchangeability And Their Applications

**Shlomo S. Sawilowsky**  
Evaluation & Research  
Wayne State University  
Editor

**Bruno D. Zumbo**  
Measurement, Evaluation, & Research  
University of British Columbia  
Associate Editor

### Assistant Editors

**Vance W. Berger**  
Biometry Research Group  
National Cancer Institute

**Todd C. Headrick**  
Educational Psychology  
Southern Illinois University  
Carbondale

**Harvey Keselman**  
Department of Psychology  
University of Manitoba

**Alan Klockers**  
Educational Psychology  
University of Washington

## Editorial Board of Journal of Modern Applied Statistical Methods

Subhash Chandra Bagui  
Department of Mathematics & Statistics  
University of West Florida

Chris Barker  
MEDTAP International  
Redwood City, CA

J. Jackson Barnette  
Community and Behavioral Health  
University of Iowa

Vincent A. R. Camara  
Department of Mathematics  
University of South Florida

Ling Chen  
Department of Statistics  
Florida International University

Christopher W. Chiu  
Test Development & Psychometric Rsch  
Law School Admission Council, PA

Jai Won Choi  
National Center for Health Statistics  
Hyattsville, MD

Rahul Dhanda  
Forest Pharmaceuticals  
New York, NY

John N. Dyer  
Dept. of Information System & Logistics  
Georgia Southern University

Matthew E. Elam  
Dept. of Industrial Engineering  
University of Alabama

Mohammed A. El-Saiedi  
Accounting, Finance, Economics &  
Statistics, Ferris State University

Carol J. Etzel  
University of Texas M. D.  
Anderson Cancer Center

Felix Famoye  
Department of Mathematics  
Central Michigan University

Barbara Foster  
Academic Computing Services, UT  
Southwestern Medical Center, Dallas

Shiva Gautam  
Department of Preventive Medicine  
Vanderbilt University

Dominique Haughton  
Mathematical Sciences Department  
Bentley College

Scott L. Hershberger  
Department of Psychology  
California State University, Long Beach

Joseph Hilbe  
Departments of Statistics/ Sociology  
Arizona State University

Peng Huang  
Dept. of Biometry & Epidemiology  
Medical University of South Carolina

Sin-Ho Jung  
Dept. of Biostatistics & Bioinformatics  
Duke University

Harry Khamis  
Statistical Consulting Center  
Wright State University

Kallappa M. Koti  
Food and Drug Administration  
Rockville, MD

Tomasz J. Kozubowski  
Department of Mathematics  
University of Nevada

Kwan R. Lee  
GlaxoSmithKline Pharmaceuticals  
Collegeville, PA

Hee-Jeong Lim  
Dept. of Math & Computer Science  
Northern Kentucky University

Devan V. Mehrotra  
Merck Research Laboratories  
Blue Bell, PA

Balgobin Nandram  
Department of Mathematical Sciences  
Worcester Polytechnic Institute

J. Sunil Rao  
Dept. of Epidemiology & Biostatistics  
Case Western Reserve University

Brent Jay Shelton  
Department of Biostatistics  
University of Alabama at Birmingham

Karan P. Singh  
University of North Texas Health  
Science Center, Fort Worth

Jianguo (Tony) Sun  
Department of Statistics  
University of Missouri, Columbia

Sheela Talwalker  
T Walker Consulting  
San Diego, CA

Joshua M. Tebbs  
Department of Statistics  
Oklahoma State University

Dimitrios D. Thomakos  
Department of Economics  
Florida International University

Justin Tobias  
Department of Economics  
University of California-Irvine

Jeffrey E. Vaks  
Beckman Coulter  
Brea, CA

Dawn M. VanLeeuwen  
Agricultural & Extension Education  
New Mexico State University

J. J. Wang  
Dept. of Advanced Educational Studies  
California State University, Bakersfield

Dongfeng Wu  
Dept. of Mathematics & Statistics  
Mississippi State University

Chengjie Xiong  
Division of Biostatistics  
Washington University in St. Louis

Andrei Yakovlev  
Biostatistics and Computational Biology  
University of Rochester

Heping Zhang  
Dept. of Epidemiology & Public Health  
Yale University

### International

Mohammed Ibrahim Ali Ageel  
Department of Mathematics  
King Khalid University, Saudi Arabia

Mohammad Fraiwan Al-Saleh  
Department of Statistics  
Yarmouk University, Irbid-Jordan

Keumhee Chough (K.C.) Carriere  
Mathematical & Statistical Sciences  
University of Alberta, Canada

Debasis Kundu  
Department of Mathematics  
Indian Institute of Technology, India

Christos Koukouvinos  
Department of Mathematics  
National Technical University, Greece

Lisa M. Lix  
Dept. of Community Health Sciences  
University of Manitoba, Canada

Fadia Nasser  
College of Education  
Tel Aviv University, Israel

Takis Papaioannou  
Statistics and Insurance Science  
University of Piraeus, Greece

Mohammad Z. Raqab  
Department of Mathematics  
University of Jordan, Jordan

Nasrollah Saebi  
School of Mathematics  
Kingston University, UK

## Journal of Modern Applied Statistical Methods

### *Invited Articles*

- 202-227     **Ronald C. Serlin**     Constructive Criticism
- 228-239     **Ralph D'Agostino,**  
**Lisa M. Sullivan**     Chronic Disease Data And Analysis: Current State Of The Field
- 240-242     **Thomas R. Knapp**     Some Reflections On Significance Testing
- 243-247     **Phillip I. Good**     Extensions Of The Concept Of Exchangeability And Their Applications

### *Regular Articles*

- 248-268     **Gail Fahoome**     Twenty Nonparametric Statistics And Their Large-Sample Approximations
- 269-280     **Vance W. Berger,**  
**Anastasia Ivanova**     Adaptive Tests For Ordered Categorical Data
- 281-287     **Rand R. Wilcox,**  
**H. J. Keselman**     Within Group Multiple Comparisons Based On Robust Measures Of Location
- 288-309     **H. J. Keselman,**  
**Rand R. Wilcox,**  
**Abdul R. Othman,**  
**Katherine Fradette**     Trimming, Transforming Statistics, And Bootstrapping: Circumventing The Biasing Effects Of Heteroscedasticity And Nonnormality
- 310-315     **Abdul R. Othman,**  
**H. J. Keselman,**  
**Rand R. Wilcox,**  
**Katherine Fradette,**  
**A. R. Padmanabhan**     A Test Of Symmetry
- 316-325     **Kimberly T. Perry,**  
**Michael R. Stoline**     A Comparison Of The D'Agostino  $S_u$  Test To The Triples Test For Testing Of Symmetry vs. Asymmetry As A Preliminary Test To Testing The Equality Of Means
- 326-332     **Mehdi Razzaghi**     On The Estimation Of Binomial Success Probability With Zero Occurrence In Sample
- 333-342     **Douglas Landsittel,**  
**Harshinder Singh,**  
**Vincent C. Arena,**  
**Stewart J. Anderson**     Null Distribution Of The Likelihood Ratio Statistic For Feed-Forward Neural Networks

- 343-353     **John N. Dyer,**  
              **B. Michael Adams,**  
              **Michael D. Conerly**     A Simulation Study Of The Impact Of Forecast Recovery For Control Charts Applied To ARMA Processes
- 354-366     **Tiffany Whittaker,**  
              **Rachel T. Fouladi,**  
              **Natasha Williams**     Determining Predictor Importance In Multiple Regression Under Varied Correlational And Distributional Conditions
- 367-378     **Pingfu Fu,**  
              **J. Sunil Rao,**  
              **Jiming Jian**     Robust Estimation Of Multivariate Failure Data With Time-Modulated Frailty
- 379-386     **Leslie A. Kalish,**  
              **Katherine Riestler,**  
              **Stuart J. Pocock**     Accounting For Non-Independent Observations In 2x2 Tables, With Application To Correcting For Family Clustering In Exposure Risk Relationship Studies
- 387-396     **Dudley L. Poston**     The Statistical Modeling Of The Fertility Of Chinese Women
- 397-404     **Pali Sen,**  
              **Mary Anderson**     Simulation Study Of Chemical Inhibition Modeling
- 405-410     **Ellen F. Sawilowsky**     Combining Quantum Mechanical Calculations And A  $\chi^2$  Fit In A Potential Energy Function For The  $\text{CO}_2 + \text{O}^+$  Reaction
- 411-419     **Joseph L. Musial,**  
              **Patrick D. Bridge,**  
              **Nicol R. Shamey**     A Longitudinal Follow-up Of Discrete Mass At Zero With Gap
- 420-427     **Moonseong Heo**     Exploration Of Distributions Of Ratio Of Partial Sum Of Sample Eigenvalues When All Population Eigenvalues Are The Same
- 428-442     **Hani M. Samawi,**  
              **Eman M. Tawalbeh**     Double Median Ranked Set Sample: Comparing To Other Double Ranked Samples For Mean And Ratio Estimators
- 443-451     **Walid Abu-Dayyeh,**  
              **Hani M. Samawi,**  
              **Lara A. Bani-Hani**     On Distribution Function Estimation Using Double Ranked Set Samples With Application
- 452-460     **Alan J. Klockars,**  
              **Tim P. Moses**     Type I Error Rates For Rank-Based Tests Of Homogeneity Of Slopes



- 461-472      **Shlomo Sawilowsky**      The Probable Difference Between Two Means When  
 $\sigma_1 \neq \sigma_2$  : The Behrens-Fisher Problem
- 473-478      **George W. Ryan,**      On The Misuse Of Confidence Intervals For Two  
**Steven Leadbetter**      Means In Testing For The Significance Of  
The Difference Between The Means

*Early Scholars*

- 479-488      **Phill Gagné,**      Best Regression Model Using Information Criteria  
**C. Mitchell Dayton**

*JMASM Algorithms & Code*

- 489-517      **Bruce R. Fay**      JMASM4: Critical Values For Four Nonparametric And/Or  
Distribution-Free Tests Of Location For Two Independent  
Samples (Fortran 90)
- 518-522      **Robert DiSario**      JMASM5: A Program For Generating All Permutations of  
 $\{1, 2, \dots, n\}$  (Visual Basic)

---

*JMASM* is an independent print and electronic journal (<http://tbf.coe.wayne.edu/jmasm>) designed to provide an outlet for the scholarly works of applied nonparametric or parametric statisticians, data analysts, researchers, classical or modern psychometricians, quantitative or qualitative evaluators, and methodologists. Work appearing in *Regular Articles*, *Brief Reports*, and *Early Scholars* are externally peer reviewed, with input from the Editorial Board; in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* are internally reviewed by the Editorial Board.

Three areas are appropriate for *JMASM*: (1) development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) development or study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods. Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Problems may arise from applied statistics and data analysis; experimental and nonexperimental research design; psychometry, testing, and measurement; and quantitative or qualitative evaluation. They should relate to the social and behavioral sciences, especially education and psychology. Applications from other traditions, such as actuarial statistics, biometrics or biostatistics, chemometrics, econometrics, environmetrics, jurimetrics, quality control, and sociometrics are welcome. Applied methods from other disciplines (e.g., astronomy, business, engineering, genetics, logic, nursing, marketing, medicine, oceanography, pharmacy, physics, political science) are acceptable if the demonstration holds promise for the social and behavioral sciences.

---

Editorial Assistant  
**Holly Atkins**

Professional Staff  
**Bruce Fay,**  
Business Manager  
**Joe Musial,**  
Marketing Director

Production Staff  
**Holly Atkins**  
**Christina Gase**  
**Bulent Ozkan**  
**Letitia Uduma**

Internet Sponsor  
**College of Education,**  
**Wayne State University**

Entire Reproductions and Imaging Solutions Internet: <a href="http://www.entire-repro.com">www.entire-repro.com</a>	248.299.8900 (Phone) 248.299.8916 (Fax)	e-mail: <a href="mailto:sales@entire-repro.com">sales@entire-repro.com</a>
--	--	---

## *INVITED ARTICLES* Constructive Criticism

Ronald C. Serlin  
University of Wisconsin-Madison



---

Attempts to attain knowledge as certified true belief have failed to circumvent Hume's injunction against induction. Theories must be viewed as unprovable, improbable, and undisprovable. The empirical basis is fallible, and yet the method of conjectures and refutations is untouched by Hume's insights. The implications for statistical methodology is that the requisite severity of testing is achieved through the use of robust procedures, whose assumptions have not been shown to be substantially violated, to test predesignated range null hypotheses. Nonparametric range null hypothesis tests need to be developed to examine whether or not effect sizes or measures of association, as well as distributional assumptions underlying the tests themselves, meet satisficing criteria.

Keywords: Probability, knowledge, satisficing, statistical methodology

---

### Introduction

In the middle of the seventeenth century, a remarkable confluence of scientists, mathematicians, and philosophers laid the foundations for the theory of probability and formulated new philosophical underpinnings for the justification of claims to knowledge. These individuals knew one another, posed problems as challenges to one another, and criticized and defended the work of one another. Although investigations in probability had been conducted for well over two hundred years before, Fermat

and Pascal were credited (by many historians of probability) with its mathematical development. Although many modern philosophical problems had been addressed by Aristotle, Socrates, and Protagoras, the interplay between probability and philosophy did not begin in earnest until the end of the seventeenth century and did not give birth to what Stigler (1986) called the infant discipline of statistics until 1900.

One reason for this fairly long dalliance is that it was not clear how the information provided by a probabilistic analysis could warrant knowledge claims, claims that at the time required justification as certain and true. Only slowly did probable knowledge get recognized as having any veracity, and this on a secondary level as opinion or belief. By the end of the eighteenth century, philosophers began to view even the possibility of acquiring certain knowledge of the real world as uncertain at best. It was only in the middle of the nineteenth century, when the philosophical focus shifted from the justification of the source of scientific knowledge to the validity of the methods of science, that the true romance between

---

Ronald C. Serlin is Professor in the Department of Educational Psychology at the University of Wisconsin-Madison. He teaches an introductory sequence in statistics, as well as courses in nonparametric statistics, multivariate statistics, and the philosophy of science and statistics. He won a University of Wisconsin teaching award, and he served two nonconcurrent terms as department chair. Email him at the following address: [rcserlin@facstaff.wisc.edu](mailto:rcserlin@facstaff.wisc.edu)

probability and philosophy blossomed in the testing of scientific theories.

This relationship continues to flourish, and the occasional disagreements are healthy, for “statistics requires a dynamic balance between its philosophical underpinnings and its practice to remain vital” (Kadane, 1976, p. 735). In order better to understand this balance and to maintain and strengthen the vitality of the applied and theoretical aspects of modern statistics, it will be helpful to examine the history of probability and its joint effort with philosophy of science. Such study will encourage researchers in statistical theory and methods to focus on problems whose solutions are essential to the continued health of the scientific enterprise, it will allow those researchers to avoid repeating mistakes of the past, and it is hoped that it will engender an appreciation for the incredible insights and magnificent oversights of our scientific forebears.

As Stigler wrote (1986), “the advances in scientific logic that took place in statistics before 1900 were to be every bit as influential as those associated with the names of Newton and Darwin” (p. 361). Indeed, even though Newton dabbled in probability theory, and Darwin’s indirect affect on statistics through his cousin, Francis Galton, is well known, less well known perhaps are Newton’s and Darwin’s influence on philosophers of science and statistics. An understanding of these kinds of mutual influences of statisticians and philosophers may help to limn modern statistics in a new yet joyously familiar way, “...a recognition, the known appearing fully itself, and more itself than one knew” (Levertov, 1961).

#### Origins of Probability Theory

According to Walker (1927), the foundations of the theory of probability were laid by Blaise Pascal and Pierre de Fermat in 1654 in response to two questions asked of Pascal by Antoine Gombauld, the Chevalier de Mere, Sieur de Baussay. As with many, if not most, scientific advances, the work of Pascal and Fermat culminated the efforts of other scientists and mathematicians that had been accruing over a period of hundreds of years. Pascal and Fermat were first brought together through the auspices

of Pierre de Carcavi and Marin Mersenne. Mathematicians, including Pierre Gassendi, Pierre de Carcavi, Gilles Roberval, Rene Descartes, and Blaise Pascal’s father, Etienne, met at Mersenne’s house once a week. Etienne introduced Blaise to the Mersenne Academy when Blaise was fourteen years old. Carcavi brought his friend Fermat, with whom he served in parliament in Toulouse, into correspondence with Mersenne and the others in 1636, and he suggested that Etienne and Roberval write to Fermat regarding their questions into methods of integration and centers of gravity. When Descartes criticized (erroneously) Fermat’s method of finding tangents, it was Etienne and Roberval who defended him. Carcavi also first put Fermat and Blaise Pascal in touch with one another (David, 1962).

One of the questions that de Mere asked, known as the problem of points, concerned the fair distribution of stakes between two players when a game they were playing was interrupted mid-contest. The problem of points had been solved more than 250 years beforehand in some works by Antonio de Mazzinghi from around 1400 (Kiernan, 2001). The first time that the problem appeared in a mathematical work, it was solved incorrectly by Pacioli in 1494 (David, 1962; Kiernan, 2001). Cardano, who offered his own solution in 1539 (four years before Copernicus published his heliocentric theory!), referred to Pacioli’s error as one that a child should recognize.

Unfortunately, Cardano’s solution was wrong. In 1556, Tartaglia again took up the problem of points, commenting that Cardano’s solution didn’t make sense. Kiernan (2001, p. 181) notes that Tartaglia’s answer was “way off”, as well. Peverone in 1558 also attempted to solve the problem and failed, but according to David (1962), M. G. Kendall called this one of the near misses of history. It was not until Pascal and Fermat discussed the problem in a series of letters during the summer of 1654 that a correct solution was again found. This time the problem of points was solved in three different ways, one by Fermat using the enumeration of all cases, one by Pascal that used the process of recursion, and a second solution by Pascal using his arithmetic triangle. (The use of a triangular array such as Pascal’s triangle to determine binomial

coefficients appeared in works by Chu Shih-chieh in 1303, Apianus in 1527, Stifel in 1545, and Tartaglia in 1556. According to David, 1962, Fermat dealt with it in 1636, which is perhaps the reason that Fisher has referred to it as Fermat's triangle.)

The second question posed by de Mere and solved by Fermat and Pascal dealt with probabilities associated with dice. He asked Pascal (and Roberval) why the probability of throwing at least one six in four rolls of a fair die was in the ratio 671 to 625, whereas the probability of obtaining at least one pair of sixes in twenty-four rolls of two dice was less than 0.5. Because the expected number of sixes rolled in four rolls of a single die is the same as the expected number of pairs of sixes in twenty-four rolls of two dice, the unequal probabilities that de Mere discovered led him, according to Pascal, to think he had found a "falsehood in the theory of numbers" and that "Arithmetic is self-contradictory" (cited in David, 1962, p. 88-89). That de Mere was able to distinguish empirically between two probabilities whose true values are 0.4914 and 0.5177, concluding that the former was less than 0.5, indicates that he was an assiduous gambler and note-taker.

Dice of reasonable quality are known to have existed since about 3000 B.C., used chiefly at the time in religious rites (David, 1962). A complete enumeration of the various outcomes on three dice appeared in a thirteenth century poem attributed to Fournival (David, 1962), and a 1477 commentary by Libri on Dante's *Divine Comedy* contains the first indication of the probabilities of various throws in a three-dice game of hazard (Todhunter, 1865). Cardano, however, possibly in concert with Ferrari, introduced in about 1526 (published posthumously in 1663) "the idea of combinations to enumerate all the elements of the fundamental probability set" and noticed that if all elements are equiprobable the ratio of favorable to total numbers of cases gives a result "in accordance with experience" (David, 1962, p. 58).

From this, David (1962) concluded that Cardano was the first mathematician to correctly calculate a theoretical probability. Unfortunately, Cardano was incorrect in his solution of what was essentially de Mere's

second question. Galileo also took up the subject of dice games and published a fragment on them in around 1620 (David, 1962). His benefactor, to whom Galileo was Mathematician to his Serenest Highness, Cosimo II of Tuscany, had posed a problem that had been solved by Cardano and that was similar to that posed by de Mere: Why, in the throwing of three dice, is the number of partitions of 9 and 10 the same, though their probability in practice was not equal, with 9 being the less probable (David, 1962)? (His Serenest Highness was almost as discerning as de Mere, being able to distinguish between probabilities of 0.116 and 0.125.)

We can see that the topics addressed by Pascal and Fermat had a long history before the summer of 1654. Nevertheless, as Todhunter (1865) commented, "neglecting the trifling hints which may be found in preceding writers, we may say that the Theory of Probability really commenced with Pascal and Fermat" (p. 20). And yet, this work was never published by either Pascal or Fermat, though both desired that it be published.

It was Christian Huygens who incorporated their work into a small tract published in 1657, the first printed work on games of chance (Walker, 1929). Huygens learned the problem of points from one of Carcavi's friends (David, 1962). After Huygens solved the problem and sent his solution to Roberval, Carcavi sent Huygens the outlines of the discussion of the problem between Fermat and Pascal, and he later sent Fermat's solution to Huygens, which turned out to be the same as Huygens'. Fermat posed even more difficult problems to Huygens, which he solved and incorporated into his tract (David, 1962). According to David (1962), if one says that "the real begetter of the calculus of probabilities is he who first put it on a sound footing" (p. 110), then one should look to Huygens, Lord of Zelem and of Zuylichem, "the scientist who first put forward in a systematic way the new propositions..., who gave the rules and who first made definitive the idea of mathematical expectation". For nearly fifty years, Huygen's work (in Latin) was the unique introduction to the theory of probability (David, 1962). Todhunter (1865) attributes a 1692 English translation of Huygens' tract to John Arbuthnot.

Newton was familiar with Huygens' writings (David, 1962). With the arrival of The Great (bubonic) Plague (1664-65), Trinity University was closed, and Newton retired to Woolsthorpe for two years to invent calculus, discover the universal law of gravitation, and prove experimentally that white light is composed of all colors. Newton's *Principia* was presented to the Royal Society in 1686 and published in 1687 (printed at Edmund Halley's expense), thirty years after Huygens published the work of Pascal and Fermat. And in 1693, Newton solved what was essentially de Mere's dice problem in response to a query by Samuel Pepys, thus revealing what David (1962) described as at least elementary knowledge of probability theory.

#### Certain Knowledge

Probability theory has clearly long been of interest to gamblers. As Bellhouse (1993) noted, "familiarity with probability theory can enhance the strategy of play." Putting the parentage of the theory aside, one must wonder, given that Pascal and Fermat's theory culminated well over one hundred years of work on probability, why the methods of probability were not beginning to be incorporated into the scientific pursuit of knowledge. David (1962) opined, "At a time when it was still possible for an able mathematician to take all knowledge for his province, moreover when dicing, and gambling with annuities, were practiced as assiduously in England as anywhere else, it is indeed strange that not only Newton but nearly the whole of the English school showed no interest in them" (p. 124-125).

David (1962) suggested that the introduction of probability into science did not come before the Renaissance "because the philosophic development which opened so many doors for the human intellect engendered a habit of mind which made impossible the construction of theoretical hypotheses from empirical data" (p. 26). One or another form of Aristotelianism was dominant at the beginning of the seventeenth century (Garber, 1995). And yet, even late into the Renaissance, during a period in which Newton seemed to have obtained hypotheses from data (despite his *hypotheses*

*non fingo* claim to the contrary), probability had yet to enter the scientific arena.

One possible reason for this late entry of probability into scientific method is that in the middle of the seventeenth century, and through the middle of the nineteenth century, knowledge was defined as certified true belief. Indeed, even Pascal claimed that he was not satisfied with the probable, seeking instead the sure (Watkins, 1978). At the heart of this epistemological view, according to Suppe (1977), was the argument that S knows that P if and only if (a) P is true, (b) S believes that P, and (c) S has adequate evidence for believing that P. From the late sixteenth through the early twentieth centuries, natural philosophers were preoccupied by systematic methods for discovering knowledge (Mulaik, 1987). In this regard, the justification clause (c) was satisfied only by finding a demonstrably incorrigible base knowledge consisting either of the intuitionist Descartes' a priori clear and distinct ideas or by the sense data of inductivists such as Bacon and Gassendi.

Greek philosophers recognized that the senses can deceive us. For example, atomists such as Democritus believed the world to be made from tiny entities known as atoms whose action on the senses cause us to experience smell and heat, for example. Yet, as the atoms have no smell or heat, the world of appearance is illusory (Mulaik, 1987). For Descartes, whom Peirce called "the father of modern philosophy" (Peirce, 1868), the broadest aspects of nature are understood by deduction from incorrigible first principles, which are grounded in pure reason (Salmon, 1966).

So committed to certainty was Descartes that in his *Discourse on Method* of 1637 he claimed as false all that was only probable. According to Cartesianism, the world is full of an infinitely divisible matter, reason dominates, and philosophy is based on his own clear and distinct perceptions (Garber, 1995). For example, as Descartes wrote in his *Meditations* (1642), "Now it is manifest by the natural light that there must at least be as much reality in the efficient and total cause as in its effect. For, pray, whence can the effect derive its reality, if not from its cause?" Salmon wonders how the intuitionist Descartes, a man who could not be certain that  $2+2=4$  or that he had hands unless he

could prove that God is not a deceiver, found it impossible to conceive of the falsity of the foregoing principle.

Descartes prepared his *Meditations* in Holland in 1640. Huygens transported it in manuscript form to Mersenne, who solicited responses from “learned men who would take the trouble to scrutinize them” (Descartes, cited in Joy, 1995, p. 431). Among those who contributed were Hobbes, Gassendi, and Mersenne, himself. According to Agassi (1975), Gassendi asked why one would deduce “I think, therefore I am?” Why not “I walk, therefore I am?” Descartes understood the point and agreed that if one walked, one necessarily existed. But he could not be sure that he walked; he could be sure that he thought, and that is why he preferred his “Cogito”. He didn’t doubt the validity of Gassendi’s inference, he only doubted the truth of the premise that he walked. (Agassi misattributed this Fifth Objection to Hobbes, who actually wrote the Third.)

Gassendi was an empiricist. For him, experience dominates, and philosophy begins with our sensations of a public world; this world is made up of atoms and a void, and he attempted to reconcile Epicurean atomism in a way that was more congenial to the Church. In rejecting Aristotelianism, he, like Descartes, adopted the mechanist philosophy’s premise that physical phenomena could be described fully in terms of matter and motion. He also believed that our senses can fool us, which caused him to formulate a kind of moderate skepticism that influenced Locke, Peirce, and others.

For other empiricists, like Bacon, the justification of scientific theory is based on its ability to explain experimental results. Until Bacon, logic as described in Aristotle’s *Organon* (Greek for “tool”) was deductive. What was needed was a method that abandoned Aristotelianism’s approach that began with hypotheses and deduced truths from them (Mulaik, 1987). Bacon introduced his inductive logic in his *Novum Organum* (Latin for “New Tool”) in 1620. According to Bacon’s doctrine (Lakatos, 1978), a discovery is scientific only if it is guided by facts through a method of induction “that would begin without hypotheses or speculations, systematically interrogate nature, and move to ever more general truths by

means of an automatic procedure or algorithms” (Mulaik, 1987, p. 273). The scientist starts by clearing his mind of theory (bias), and nature will then make itself known. For Bacon, science is an experimental enterprise through which one investigates phenomena in controlled circumstances. Bacon’s method of eliminative induction includes the logical insight that affirming instances do not provide evidence for inductive generalizations, whereas negative instances do provide disconfirming evidence (Mulaik, 1987). Bacon, apocryphally, died of pneumonia that developed while he was investigating refrigeration by stuffing a chicken with snow.

Although Bacon’s *Novum Organum* of 1620 preceded Descartes’ *Discourse on Method* by seventeen years, Descartes’ philosophy was dominant at the time of Newton’s *Principia*. According to the justificationist standards of the day, then, Newton’s theory was non-knowledge (Lakatos, 1978). Newton’s theory was not proved in the Cartesian sense, because it was not derived from Cartesian metaphysics. Newton instead proposed that propositions required only an empirical-experimental and not a rational-metaphysical proof (Lakatos, 1978). Because of the extraordinary success of Newton’s theory, “for 200 years after Newton no one could advocate the use of hypotheses without an uneasy backward glance” (Medawar, 1974). This, despite inductivism having suffered what should have been severe setbacks at the hands of Locke, Hume and Kant.

#### Probable Knowledge

The beauty and power of Newton’s mathematical approach to physics clearly had an effect on John Arbuthnot, who wrote in 1692, “There are very few things which we know; which are not capable of being reduc’d to a Mathematical Reasoning; and when they cannot, it’s a sign our Knowledge of them is very small and confus’d” (Stigler, 1986, p. 1). Arbuthnot implemented a binomial test in 1710 to examine “the constant regularity observ’d in the births of both sexes,” (Stigler, 1986, p. 225), and he is often credited with publishing the first statistical test. Fisher, however, attributed the first published significance test to de Moivre in 1718, and Barnard stipulated that the first published

test was due to Daniel Bernoulli in 1734 (Bennett, 1990, p. 23-26). Regardless of which test is deemed to have been the first, it is clear that the eighteenth century held promise for great discoveries in probability and statistics. Some of the early discoveries in probability and statistics were important to philosophers, as well. Jacob Bernoulli developed the theory of permutations and combinations and contributed the weak law of large numbers, the theorem that with an increasing number of observations, the probability increases that an estimator will lie within any specified distance of the true value.

According to Stigler (1986), at least five Bernoullis worked on probability, writing "So large is the set of Bernoullis that chance alone may have made it inevitable that a Bernoulli should be designated father of the quantification of uncertainty" (Stigler, 1986, p. 63). Jacob Bernoulli and philosopher Gottfried Leibniz are known to have composed twenty-one letters to one another, although one may not have been sent (Sylla, 1998). Leibniz may have first learned of Jacob's work in probability from Jacob's brother, Johann, with whom Jacob was not speaking. In a letter written in 1697, Leibniz spoke of the "need for establishing on firm foundations an art of measuring degrees of proofs" (Sylla, 1998, p.48). And after the publication in 1713 of Jacob Bernoulli's *Ars Conjectandi*, accomplished eight years posthumously by his nephew Nicholas because of the rift between brothers, Leibniz noted that the probabilities of obtaining an 11 and a 12 in rolling two dice are equal.

John Locke is considered to be the father of British empiricism, and he is perhaps the first major philosopher to discuss probable knowledge as a somewhat tenable, "second-rate way of becoming cognitively aware of the nature of the world" (Owen, 1993, p. 38). For Locke, probable knowledge is faith or opinion. Owen noted that Locke and other non-Cartesians stood at a junction between the old and new ways of looking at the world. Locke's account "recognizes the limitations of knowledge, rather traditionally conceived, but looks ahead in allowing its rational supplementation by probable conjectures" (Owen, 1993, p. 39).

In his 1690 *An Essay Concerning Human Understanding*, Locke sought to support

Bacon's empiricism by arguing that knowledge can not have a component based on innate ideas. He argued that if knowledge is not received through the senses, then the mind at birth must have some kind of intellectual ability, at least in applying the concepts of logic (Clark, 1957). Instead, he felt that a person enters the world with a mind that is a blank slate. There are only two sources of ideas, sensation and reflection. For Locke, complex ideas are formed out of the simple ones entering the mind through the mental activities of compounding, abstracting, and relating. By a method of analysis, Locke was able to trace back from complex ideas to the simple ones out of which they arose, but he could not find the simple idea from which the concept of substance came (Mulaik, 1987). Because of this, and because he argued that the certain qualities of objects, such as color and odor, exist only in the mind and are not representative of reality, we can not be certain that any of our ideas are representative of reality.

The case for the demise of inductivism was made well and irremediably in David Hume's *Enquiry concerning the human understanding* of 1748. Hume's objections to induction can be variously phrased. According to Harris (1992), Hume concluded that it is impossible to justify epistemologically that unobserved cases will resemble observed cases in some crucial respect. Because of this, neither certain nor probable knowledge can be justified.

Reichenbach (1951) discussed two theses put forward by Hume. In the first thesis, Hume makes clear the nonanalytic nature of induction by pointing out that we can very well imagine the contrary of the inductive conclusion. The possibility of a false conclusion in combination with a true premise proves that the inductive inference does not carry a logical necessity with it. Hume's second thesis is that induction cannot be justified by reference to experience--the inference with which we want to justify induction is itself an inductive inference (we believe in induction because induction has so far been successful), and so we are caught in circularity. Russell (1945, p. 672) stated Hume's conclusion as, "We cannot help believing, but no belief can be grounded in reason." Of Hume's conclusion, Russell (1945) exclaimed, "It is therefore important to discover whether there is

any answer to Hume within the framework of a philosophy that is wholly or mainly empirical. If not, there is no intellectual difference between sanity and insanity. The lunatic who believes that he is a poached egg is to be condemned solely on the ground that he is in a minority" (p. 673). It would seem that as of 1748, unless arguments could be mounted against Hume's attack, inductivism was dead. Yet, it lived on, because of the success of Newton's theory.

Expanding on the work of Jacob and Nicholas Bernoulli, De Moivre published the first appearance of the normal curve in 1733 (Stigler, 1986). And in 1763, Bayes' Theorem was published posthumously by Richard Price, who presented it to the Royal Society. Fisher (1956) thought Bayes was reluctant to publish his work because Bayes felt that his postulating a uniform prior distribution might be considered disputable. Price, according to Gillies (1993), was strongly influenced by Hume's criticisms of induction and thought that Bayes' Theorem could be used to resolve the problems raised by Hume by making generalizations probable, rather than certain (this despite Hume's injunction against such a possibility).

#### Synthetic *a priori* Knowledge

The first major intuitionist response to Hume's empiricist attack was due to Kant, who wrote *Critique of pure reason* in 1781, according to Reichenbach (1951), "with the intention of saving scientific knowledge from the annihilating consequences of Hume's criticism." Kant, who in his preface to the *Critique* compared his work to that of Copernicus, made clear two distinctions among types of propositions. First, he distinguished between analytic propositions (true virtually by definition, such as the statement "All bachelors are unmarried") and synthetic propositions (those that inform us about a fact, such as observations, and add to our knowledge). Second, he distinguished between *a priori* propositions, those which have a basis other than experience, and *a posteriori* (or empirical) propositions, needing observational evidence to determine their truth. He posited that objects conform to the conditions set forth by the mind, that whereas the senses provide the subject matter, the mind imposes the form of thought.

Rather than the mind being a Baconian blank slate, Kant specified what he called the categories of thought as the *a priori* equipment for thinking. He felt that by showing that the axioms of Euclidean geometry were synthetic and yet known *a priori*, he could establish the incorrigible basis that justified Suppe's clause (c) mentioned earlier. It would seem, then, that at this point, intuitionism held the upper hand, due to Hume's crushing blow against inductivism and to Kant's intuitionist argument that Euclidean geometry was synthetic and yet known *a priori*.

The nineteenth century saw major upheavals in science and philosophy. As described by Reichenbach (1951), "Ever since the death of Kant in 1804 science has gone through a development, gradual at first and rapidly increasing in tempo, in which it abandoned all absolute truths and preconceived ideas." Lagrange introduced the method of least squares in 1805, and in 1809 Gauss addressed the same problem but couched it in probabilistic terms (he also claimed priority for the method of least squares, claiming he had used it since 1795 - Stigler, 1986).

Laplace contributed the central limit theorem in 1810, inverse probability and the principle of insufficient reason in 1812. His definition of probability was as a state of mind (Fisher, 1956; Epstein, 1977), whereas Bayes seems to have used a frequentist definition (Fisher, 1956). The definition of probability as the limit of a frequency was due to Poisson in 1837. According to Epstein (1977), the theory of probability is more indebted to Laplace than to any other mathematician; indeed, Stigler (1986, p. 122) claims that Laplace's work brought about "a truly Copernican revolution in statistical concept." The Gauss-Laplace synthesis brought together two lines - the combination of observations and the use of probability to make inferences - into a coherent whole that was widely disseminated through the middle of the century (Stigler, 1986).

But Gauss, along with Bolyai and Nikolai Ivanovich Lobachevsky, called the Copernicus of geometry by English mathematician William Clifford (Bell, 1937), made a discovery that had far greater philosophical import - the discovery of



non-Euclidean geometry. Lobachevsky's publication appeared in 1829-30 and Bolyai's in 1832. Gauss claimed to have obtained similar results earlier but did not publish because, according to Gillies (1993, p. 80), "he was 'afraid of the clamour of the Boeotians.' Boeotia was a region of ancient Greece whose inhabitants were considered by the Athenians to be stupid and uncultured" (p. 80). The arrival of non-Euclidean geometry showed that Kant's implication that humans could never conceive of non-Euclidean geometries was untenable. Despite this, Kant's impact was strong and lasting.

#### Descriptive Knowledge

Burt (1924) saw elements of positivism in Galileo's work, and Burt cited Brewster's claim that Newton was the first great positivist. The founder of positivism in its 19th century form was Auguste Comte. Comte's *Cours de philosophie Positive* was completed in 1842. Comte is also known as the founder of sociology. Positivism was Comte's response to the upheavals in society and to Laplace's "scientifically reasoned deterministic interpretation of the universe" (Epstein, 1977, p.7). It was Comte's hope that science could be turned into a religion, "in which the great philosophers and scientists took the place of the Christian saints, and an organized devotion to the cause of humanity was substituted for the worship of God" (Fuller, 1938, p. 384). According to Comte, there are three stages in the history of thought: 1) a theological stage, explaining the universe in terms of the purposes of deities; 2) a metaphysical stage, explaining in terms of abstract principles which are personified; and 3) a scientific stage, in which uniformities in nature are described without reading any evidence of purpose or design or consciousness into them. The meaning of terms are referred to what is found in experience.

Positivists eschew metaphysics and refrain from explanation in physics. Science organizes knowledge using laws that are merely descriptions, approximate at that, of the patterns in which phenomena occur, and science gives us the power of prediction. Bradley (1971) paraphrased Martineau in saying it is strange

that something so negative should be called positivism.

Fortunately, although an actual Religion of Positivism was started, with priests, rituals, and baptisms, most of Comte's excesses in this direction were ignored. Comte's positivist heir was physicist Ernst Mach, who was ecumenical in his influences, including Hume, Kant, and Darwin (Cohen, 1970, p. 127). According to Cohen (1970), Mach "apparently succeeded in combining a Kantian appreciation of the active, even constitutive, role of the mind in generating science with a scientific, which is to say, empirical-biological, theory of the origins and functions of the mental life" (p. 156). For Mach, "not knowledge attained, but the method of attaining it, could be certified" (Cohen, 1970, p.129).

Mach, like Comte, was an instrumentalist and felt that laws were mere descriptions of nature. Mach, however, did not completely do away with theories (as opposed to laws), as long as they were testable. Mach's positivism differs from Comte's in that nothing was "more foreign to Mach than the tendency towards absolutism which finally disfigured both the philosophical and the human image of Comte" (von Mises, 1970, p. 266). Even by the turn of the twentieth century, physicists such as Plank and Einstein, although influenced greatly by Mach early on, began to turn against positivism.

#### Conjectural Knowledge

William Whewell, who coined the word 'scientist' (as well as 'anode' and 'cathode' for Faraday and the words 'physicist', 'eocene', 'miocene', and 'pliocene' - Medawar, 1974) upon the request of the poet Samuel Taylor Coleridge in 1833, tried to reformulate the problems of the philosophy of science in a Kantian way (Wettersten, 1993), while not relying on Kant's fixed a priori categories. He attempted to "explain the facts of the growth and stability of science without appeal to induction, which he saw to be useless" (Wettersten, 1993, p. 482). In his *Novum Organum Renovatum* of 1858, Whewell considered induction to be "the representation of facts with principles" (Wettersten, 1993, p. 497), a notion that will be seen in the pragmatist philosophy of Charles

Sanders Peirce, and not the Baconian induction from facts to generalizations. He showed that neither empiricism nor intuitionism, including Kant's, could account for the growth of scientific knowledge; instead, both experience and intuition were needed. He gave importance to independent tests and to new predictions, and he claimed that science needs guesses (Medawar, 1974 noted that Whewell also used the phrase 'felicitous strokes of inventive talent' when a more formal phrase than 'happy guesses' was required.) As Medawar (1974, p. 281) explained, "To say that Einstein formulated a theory of relativity by guesswork is on all fours with saying that Wordsworth wrote rhymes and Mozart tuneful music. It is cheeky where something grave is called for to explain how scientists discover true principles." According to Wettersten (1993, p. 506), Whewell's theory makes clear that "even if we start with poor guesses and treat them critically we can come to the truth: there are many paths to the truth but only one goal'. We see then that Whewell's approach is essentially deductivist and that the process consists above all in criticism. In this, Whewell is a direct predecessor to Karl Popper's philosophy of conjectures and refutations (Wettersten, 1992).

According to Reichenbach (1951), "the turning point in the history of logic was the middle of the nineteenth century, when mathematicians like Boole and de Morgan undertook to set forth the principles of logic in a symbolic language." Peirce, a mathematician and logician by training, carried on this work. It was not until Boole, DeMorgan, and Peirce mathematically overhauled traditional formal logic that the logic of probability was put on a more scientifically useful basis (Wiener, 1972). That Peirce was a frequentist could have been due to Boole's strong criticism in 1854 of the postulate of which Bayes was so chary. Like Whewell, Peirce was heavily influenced by Kant. He claimed that he read Kant's *Critique of Pure Reason* two hours per day for three years, and he named his philosophy 'pragmatism' in honor of Kant, whom he called The Philosopher. He did not use the term practicalism, because in Kant pragmatism and practicalism are virtually polar opposites (Buchler, 1939).

Pragmatic means empirical or experimental, whereas Kant's notion of practical laws are given purely a priori. Indeed, so often were these terms misunderstood that Peirce threatened to call his philosophy pragmatism, a term he felt was so ugly that it wouldn't be kidnapped.

According to Wiener (1972), the great difference between the American pragmatists and Kant is their denial that over and above contingent pragmatic belief are the purely rational, necessary, and absolute ideas of Kant's transcendental philosophy. The purpose of inquiry, wrote Peirce, is to enable us to pass from a state of doubt to a state of belief. Despite his high regard for Kant, Peirce's philosophy differed from that of Kant. For example, whereas Kant considered mathematics to be synthetic and yet true a priori, Peirce held that mathematics and logic are not synthetic (Buchler, 1939).

He also provided his own version of Kant's categories, writing of them that in making their character unchangeable, Kant was hostile to the spirit of empiricism. Because of the constant nature of Kant's categories, Kant's epistemology formed a closed system. But Peirce, having the benefit of Darwin's *Origin of Species* of 1859, provides an adaptive mechanism behind his categories. Peirce attempted to convert the Darwinian ideas of chance variation and natural selection into the idea of an evolution of the mind by means of a logical competition among thoughts, which eliminates ideas not fit to stand for the truth fated to be discovered by those who investigate. It was the nonevolutionary character of the old forms of a static empiricism and a rigid *a priori* intuitionism that engaged the pragmatists.

Peirce was a fallibilist, extending the views of Gassendi and Locke in a most thorough way. "I will not," he wrote, "admit that we know anything with *absolute certainty*. It is possible that twice two is not four" (Peirce, 1958, p. 64). Although he felt that the notion of certain knowledge is absurd for a variety of reasons, there were two main reasons underpinning his fallibilism. First, all claims to knowledge are criticizable and only held conditionally, for there is no ultimate inductivist or empiricist basis that can stop the respective infinite regress in the

justification of the claims. And second, he felt that no theory was true, able to satisfy all features of the facts. In terms of Newton's law of gravity, he pointed out that if, instead of inverse square attraction, the exponent of the distance between bodies was 2.000001, there would only be a minor consequence observable in the orbits of the planets, resulting in only slight discrepancies in estimated planet masses (Peirce, 1958).

Peirce (1878) classified all inference as either deductive (or analytic) or synthetic, which he subdivided into induction and hypothesis. (One difficulty encountered in reading Peirce results from his using 'hypothesis', 'retroduction', and 'abduction' for the same synthetic inference. In addition, Peirce delineated several types of induction.) Deduction is a syllogism in which the truth of a rule and a case is transmitted to the result, and conversely from the falsity of the conclusion, the falsity of the premise follows. In induction, we infer from a number of cases that the same thing is true of a whole class. Peirce showed that an induction is the inverse of a deductive syllogism, so that from the case and the result, the rule is inferred. As an example (Peirce, 1878), from the deduction:

Rule: All the beans in the bag were white.

Case: These beans were in the bag.

Result: These beans are white.

we can obtain the induction:

Case: These beans were in the bag.

Result: These beans are white.

Rule: All the beans in the bag were white.

Hypothesis infers the case from the rule and the result:

Rule: All the beans from this bag are white.

Result: These beans are white.

Case: These beans are from this bag.

Peirce described the scientific method in terms of these three modes of inference in the following way (Peirce, 1958):

Accepting the conclusion that an explanation is needed when facts contrary to what we should expect emerge, it follows that the explanation must be such a proposition as would lead to the prediction of the observed facts

A hypothesis then, has to be adopted, which is likely in itself, and renders the facts likely. This step of adopting a hypothesis as being suggested by the facts, is what I call *abduction*.

[T]he first thing that will be done, as soon as a hypothesis has been adopted, will be to trace out its necessary and probable experiential consequences. This step is *deduction*. (p. 122).

An abduction for Peirce is an explanation.

The third step in the process involves induction (Peirce, 1958):

Having...drawn from a hypothesis predictions...we proceed to test the hypothesis by making the experiments and comparing those predictions with the actual results of the experiment.

This sort of inference it is, from experiments testing predictions based on a hypothesis, that is alone properly entitled to be called *induction*.

Induction...is not justified by any relation between the facts stated in the premisses and the fact stated in the conclusion...But the justification of its conclusion is that that conclusion is reached by a method which, steadily persisted in, must lead to true knowledge in the long run. (p. 124-125)

Peirce distinguished two major types of valid induction (there is actually a third type that Peirce called the Pooh-pooh argument, but enough said). The first, quantitative induction, involves the ascertainment of a ratio in the population from samples. Through this type of induction, we can attain moral certainty of the population value, by which Peirce means a probability of 1 based on Bernoulli's results concerning the probability that the sample value lies within certain limits of the population value. "Of course," he wrote, "there is a difference between probability 1 and absolute certainty" (Peirce, 1958, p. 131). The second type of induction Peirce called qualitative induction, from which the most that can be said is that there is no reason yet for giving up the hypothesis. Of this second type, Peirce (1958) wrote, "the only justification for this would be that it is the result of a method that persisted in must eventually correct any error that it leads us into" (p. 134).

Peirce claimed for induction a trustworthiness because of the manner of proceeding (Buchler, 1939). The concept of a probable argument referred to a class of arguments, and an induction belongs to the class of all inductions. Saying an induction was probable meant that the majority of inductions were successful. "[T]hat real and sensible difference between one degree of probability and another...is that in the frequent employment of two different modes of inference, one will carry truth with it oftener than the other" (Peirce, 1878).

Neither qualitative nor quantitative induction and the associated probabilities of success involves the probability that a generalization itself is true. According to Buchler (1939), "After 1883 Peirce does not even regard induction as 'probable'...but rather as not probable at all" (p. 251). Peirce said that talking about the probability of a law was nonsense, as if universes were as plentiful as blackberries, and we could pick one. This later view reflects Peirce's distinction between two types of probability, the empirical probability associated with ratios or with the class of inductions and what Peirce called conceptualistic probability that is not strictly a

probability, but is instead only a sense of probability (Buchler, 1939).

As with Whewell, Peirce emphasized that potential explanatory hypotheses are formulated as guesses. For Peirce, as with Mach, the force of scientific reason lies in its methods. "[T]he *method of methods*, is the true and worthy idea of the science" (Peirce, 1958, p. 44). Science is rational, according to Peirce (1958, p. 49), where "...'rational' means essentially self-criticizing, self-controlling and self-controlled, and therefore open to incessant question." And rather than leading to the probability that the inductive inference itself is true, the ability to draw valid conclusions lies with the probability of correctness of its inductive method, "the relative frequency with which this class of inferences is found to yield true conclusions" (Buchler, 1939, p. 233).

#### Unprovable and Improbable Knowledge

By the end of the nineteenth century, the philosophical focus was on American Pragmatism and Machian positivism. Both Galton and Pearson were Machian instrumentalists, which would at least partly explain Pearson's emphasis on fitting data to his own system of curves. The continuation of Mach's doctrines fell to the logical empiricists. The response of Russell and the Vienna Circle philosophers was to search for an empirical basis and an inductive logic. Realizing that justifying an inductive principle on the basis of observation would lead to an infinite regress - to justify it would require inductive inferences - Russell advocated accepting the principle of induction on the ground of its intrinsic evidence (Gillies, 1993), that is, on an a priori basis. But even if we accepted *a priorism* as a justification of an inductive principle, the positivists' search for an empirical basis was doomed to failure, as shown by Duhem, who advanced two theses against inductivism. One of these, afterwards to become known as the Duhem-Quine thesis, will be discussed later.

The other thesis shows that all observations are theory-laden. According to Agassi (1983), the claim that empirical evidence has a theoretical bias was recognized by Bacon and Galileo; if one has a theory, it biases perception. This led to Bacon's request that

scientists first make observations with no theory in mind. Galileo realized, of course, that this would result in “just a heap of observations” (Agassi, 1983, p. 10), and he was convinced that geometry, based on *a priori* intuitions, must precede facts. This led to Kant’s argument against empiricism, and Whewell, influenced by Kant, deduced that all data are interpreted, either on the basis of theory or of *a priori* intuitions. Therefore, trying to prove a theory inductively ultimately requires proving a theory from a theory, which is impossible. All one could conclude on this basis is that the theories involved are consistent. Thus, the theory-ladenness of observations meant that theories could no longer be hoped to be proved from an incorrigible basis.

It was still felt, however, that although theories may not be provable, they still could be disproved, or falsified, a view that flies in the face of the Duhem’s second thesis, which states that an experiment can never condemn an isolated hypothesis but only a whole theoretical group. Underpinning this thesis is the realization that no theory can specify any observable consequences. Rather, it requires the conjunction of the theory, initial conditions, and auxiliary hypotheses. Thus, there can not be such a thing as a crucial experiment, on the basis of which a theory is falsified and dropped, because an observation contrary to prediction can only condemn the collective and not any individual part. Quine (1951) concluded that any statement can be held to be true, if we make enough adjustments elsewhere in the system. Thus, not only did the theory-ladenness of observations make theories unprovable, the Duhem-Quine thesis makes them undisprovable. So positivists had to fall back on the hope that theories could at least be shown to be probable.

Neyman and Pearson (1933) and Fisher (1935) approached these issues from different perspectives, and certainly different from the probabilist approach of Jeffreys (1939). For probabilists, theories have different degrees of probability (Lakatos, 1978). Scientific honesty then consists in uttering only highly probable theories, or the probability in light of the evidence. But Ritchie (1926) showed that the probability of any inductive generalization is zero, and Lakatos (1978) points out that in the

early 1940’s, Carnap found that the degree of confirmation of all genuinely universal propositions was zero. So not only can no theory be proved or disproved with certainty, but theories are also equally improbable. This, then, was finally the end of positivism.

#### Criticism and Knowledge

Popper, in his *Logic der Forshung* in 1934 (Popper, 1959), attempted to address the issues that have been raised, especially Hume’s skepticism, the theory-ladenness of observations, and the inability to condemn a hypothesis in isolation. In his solution, we can see much of what was good in Hume, Kant, Mach, and especially Whewell and Peirce. Popper’s view of knowledge is fallibilist, as was Peirce’s, and for him method is fallible as well, as distinguished from Mach’s view that method was certain. Indeed, Peirce’s overall view of the inductive process is virtually indistinguishable from the conjecture-and-refutation model advocated by Popper (Wiener, 1972). Popper (1962) claimed that his method of conjectures and refutations had its origins in the writings of Kant. Popper never questioned Hume’s indictment of induction; instead, he insisted there was no problem. Instead of an inductive principle, Popper advanced “the theory of the deductive method of testing, or as the view that a hypothesis can only be empirically tested--and only after it has been advanced” (Popper, 1959, p. 30).

Musgrave (1993) described Popper’s solution to the problem of induction in the following way. Popper, he said, rejected the assumption that an ampliative hypothesis is reasonable if, and only if, it is justified by the evidence, if, and only if, the evidence shows it to be true or probably true. In this, it is not clear whether justifying beliefs refers to justifying the things we believe or providing a warrant for our believing those things. According to the classical argument, we are justified in believing something if, and only if, we can show it to be true or at least show it to be more likely true than not. Popper rejected this assumption, allowing him to endorse Hume’s inductive skepticism while rejecting his irrationalism. To get from the skeptical thesis to the irrationalist thesis you also must assume that a belief is

reasonable if and only if it is justified. Popper rejected this also.

In Musgrave's (1993) view, Popper affirmed that some evidence-transcending beliefs are reasonable. The central claim of Popper's approach, said Musgrave (1993), is that an evidence-transcending belief is reasonable if, and only if, it has withstood criticism, including, where appropriate, attempts to refute it by appeal to evidence. When a prediction is falsified we will say that what we predicted was wrong, not that it was unreasonable to have predicted it. For any reasonable theory of reasonable belief, according to Musgrave (1993), must make room for reasonable beliefs in untruths. In short, Hume's criticism of induction applied to the search for a warrant for our beliefs, whereas in Musgrave's view, it does not apply to obtaining a warrant for our act of believing.

By contrast, according to the pancritical rationalism of Bartley (1984) and the comprehensively critical rationalism of Miller (2002), reflecting and extending the philosophy of Peirce and Popper, "neither beliefs nor acts of belief, nor decisions, nor even preferences, are reasonable or rational except in the sense that they are reached by procedures or methods that are reasonable or rational...Still less are beliefs, or decisions, or preferences ever justified" (Miller, 2002, p. 81). According to Miller (1982), the major difference between Popper's falsificationism and the justificationist philosophy of others is methodological, not epistemological.

Virtually all modern philosophers of science agree that certain knowledge can not be attained. Popper was the first to say outright that the attempt to attain certainty should not even be made. Miller (1982) pointed out that for justificationists, a hypothesis has to be confirmed, perhaps inductively, before it is admitted to science, and if it fails the tests, or is disconfirmed, or not confirmed at all, it is excluded from science. For Descartes, ideas that can not be justified by being reduced to clear and distinct ideas should be rejected, and anything that is accepted must be justified in this way. For Hume, any idea that can be justified by being derived from experience, the empiricist's only source of knowledge, should be accepted,

and any idea that can not should be rejected (Bartley, 1984).

For Popper, as with Peirce, a hypothesis is tested only after it is admitted by being conjectured. There is a policy of "open admission", restricted only by the requirement that no hypothesis be admitted without there being some way to test it (Miller, 1982, p. 22). If the hypothesis passes a test, nothing happens, whereas if it fails a test, it is expelled. Because of the open admission policy, "it is of the greatest importance that the expulsion procedures should be brought into play at every possible opportunity...If we are seriously searching for the truth, we should submit any hypothesis proposed to the most searching barrage of criticism, in the hope that if it is false it will reveal itself as false" (Miller, 1982, p.23).

#### Criticism

One objection that could be raised regarding the critical rationalist methodology concerns the use of logic in a rational approach to science. Surely, this line of thinking would go, the principles of logic must be assumed to be true on an *a priori* basis. Are we not committed to an un-revisable logic, because logic itself can not be used to criticize logic? It is true that "critical argument...cannot be carried on without some system of logic. You cannot in this sense abandon logic and remain a rationalist" (Miller, 1994, p. 91). But the system of logic one uses can be criticized if the logical rules consistently lead to errors. Miller (1994) gives the example of a program written in FORTRAN that can be used to test the correctness of an operating system, even though the operating system is presupposed. Miller (1994) noted that it is "logic itself" (p. 91) that is supposedly assumed to be beyond criticism by critical rationalism. Yet, logic is involved in the critical argument in a particular formulation, at a minimum usually involving the principle of noncontradiction and the law of excluded middle, which might be right or wrong, and not in an unformulated way as logic itself. And whatever the particular formulation, it can certainly be criticized.

Does not the approach presuppose an inductive principle, such as the uniformity of nature or that the future will resemble the past, at least as far as specifying that we expect that

the laws we've discovered should work in the future? As Miller (1982) pointed out, "In order to provide genuinely interesting knowledge of the world inductivism needs to assume that there is some order and regularity in the world, whilst falsificationism requires only that there is some order and regularity in the world—but it does not need to make any sort of assumption to this effect" (p. 33). Miller went on to note that if there were no regularity, falsificationism would yield little, except the conjecture that there is no regularity. Hypotheses propose order, but if there is none, none will be found. They do not presuppose it.

As regards the reliability of a theory, no theory is reliable, in that Hume showed that without an inductive principle such as that the future will resemble the past, there is no logical way to infer that the theory will work in the future (or that it will fail). But if a theory is conjectured and stands up to severe testing, then it has not been discredited (a term used to emphasize the tentative nature of falsifications), and it may be tentatively classified as true; and one can *deduce* from the conjecture that various predictions will hold without relying on the uniformity of nature. As Miller (1980) wrote, "Whatever one calls them, Hume's problem simply does not arise for guesses" (p. 123). But, the issue might be pursued, if theories are unreliable, then why should any decisions be based on them?

Again, it seems rational to base a decision on a theory that has stood up to severe testing instead of one that has failed a severe test. As Miller (2002) pointed out, if one wants to avoid bad outcomes tomorrow, he can cross his fingers or he can try to be rational today. This does not mean, of course, that we can not hope that our favorite theories will continue to stand up to severe criticism. Radnitzky (1982) explained, "we have a subjective belief that the regularities described by a highly corroborated theory will also hold in the future. But this subjective belief is not granted any methodological significance" (p. 74).

Finally, the question arises as to how one could base a rejection of theory on the basis of experience if all basic statements are tentative. In this regard, Popper (1985) pointed to the well-known asymmetry between

corroboration and rejection, namely that no matter how many confirmatory observations are observed, a theory can never be proved, whereas a single disconfirmatory observation can falsify (tentatively) a theory. Thus, as regards the observational basis, "No matter whether they are true or whether they are false, a universal law may not be derived from them. However if we assume that they are true the universal law may be falsified by them" (Popper, 1985, p. 185). Here the basic statements are conjectured to be true and are severely tested. "No falsification is conclusive," Miller (1982) wrote, "if only because all test statements are themselves fallible and open to dispute. But it would be incorrect to conclude from this that no hypothesis can be properly falsified... [T]hat a falsification has not been done conclusively does not mean that it has not been done correctly" (p. 24). The important thing about basic statements, Miller (1982) pointed out, is that they should be true. If there is doubt about a basic statement, it is rational to test it. It is not enough simply to doubt, because doubt is not the same thing as criticism.

#### Gambling with Nature

The philosophical underpinnings of the demand for severity in testing hypotheses has been discussed and codified by Mayo (1996). "What are needed," she wrote (Mayo, 1996), are arguments that *H* is correct, that experimental outcomes will very frequently be in accordance with what *H* predicts—that *H* will very frequently succeed... We obtain such experimental knowledge by making use of probabilities—not of hypotheses but probabilistic characteristics of experimental testing methods (e.g., their reliability or severity)" (p. 122).

Mayo (1996) explained, "The control of error probabilities has fundamental uses in learning contexts. The link between controlling error probabilities and experimental learning comes by way of the link between error probabilities and severity. The ability to provide methods whose actual error probabilities will be close to those specified by a formal statistical model, I believe, is the key to achieving experimental knowledge" (p. 411).

Mayo seemed to concur with Peirce in this, including Peirce's focus on verification.

Yet, as we have seen, inductive support is not possible. Miller (1982) described the task of empirical science as separating as best it can true statements about the world from false ones, and to retain the true ones. The mission, of course, is to classify, and not certify, truths. Scientific conjectures are “hopelessly fallible, hopelessly improbable, hopelessly unlikely to be true” (Miller, 1982, p. 20). And yet, the conjectural nature of our hypotheses makes them ready to be shown to be wrong. In so doing, we must strictly control the rate at which we make errors in order to ensure a desired level of severity. This imposition of severe testing is a methodological one (Miller, 1982), and it is consistent with both Peirce’s philosophical views and with Neyman’s (1957) philosophy of inductive behavior.

Neyman (1957) wrote that the concluding phase of scientific research, often labeled inductive reasoning, involves mental processes that are very different from those involved in proving a theorem. Instead of inductive reasoning, which may be considered a misnomer, Neyman preferred the phrase inductive behavior. Neyman pointed out that theories are models of natural phenomena, that is (Neyman, 1957, p. 8)

A model is a set of invented assumptions regarding invented entities such that, if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypotheses constituting the model are expected to agree with observations.

In describing the concluding phase, which he pointed out was frequently described as induction, he felt that the constituent processes were of three types (Neyman, 1957, p. 10). First, the visualization of several possible sets of hypotheses relevant to the phenomenon, second deductions from these sets of hypotheses, and third an “act of will or a decision to take a particular action, perhaps to assume a particular attitude towards the various sets of hypotheses.” We need to specify in advance the desired properties of our decision procedure and try to determine the decision rule that has these properties. Given that the hypothesized model is adequate, probability calculations are used to “tell us how frequently the given rule will prescribe any of the actions contemplated”

(Neyman, 1957, p. 18). The mental processes involved in the third step, according to Neyman, amount to taking a calculated risk.

Levi (1980) commented on the connection between Peirce’s approach to induction and the Neyman and Pearson theory of hypothesis testing: “Peirce’s inductions are inferences according to rules specified in advance of drawing the inferences where the properties of the rules which make the inferences good ones concern the probability of success in using the rules. These are features of the rules which followers of the Neyman-Pearson approach to confidence interval estimation would insist on” (p. 138). Peirce’s call for predesignation is echoed in Pearson’s (1936) insight that “to base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the hypothesis is true” (p. 317). Mayo (1993), in drawing out the common philosophical underpinnings of the Peirce and the Neyman-Pearson methodologies, noted that Birnbaum and Armitage showed that violating predesignation permits tests which can be wrong with extremely high probability.

It may be illustrative to view the appropriate use of statistical methods in the course of taking Neyman’s calculated risk as a system to use, similar say to a system for playing blackjack, while “gambling with truth” (Levi, 1967) in what Milnor (1954) called “games against nature.” In a sense, probability theory is returned to its roots. If the game against nature is to be played, it seems only rational to adopt a system that is known to yield a particular advantageous probability of winning.

In blackjack, even the best systems yield an overall probability of winning of 0.51 or so (Epstein, 1977), so a player must follow a system rigorously or the chances of winning will be reduced, if not reversed. The system is not totally rigid, in that each decision is based on the available information at the time the decision is to be made, but this adaptive decision-making scheme is figured into the overall winning probability, which is known in advance. The player must be steered against following



intuition or building up superstitions. If a high card is needed, and if the cards so far observed indicate that there is a sufficient proportion of high cards left in the deck to require the player to request a card, the decision should not be influenced by having seen the previous three players receive high cards; nor by the memory that taking a card in a previous similar circumstance led to a losing hand; nor by the feeling that the queen of diamonds is an unlucky card.

Analogously, if prior theoretical or empirical information led on the basis of superior power in a three-group design to the choice of Fisher's (1935) Least Significant Difference (LSD) method of planned comparisons, then that must be the procedure that is carried out. There will be losing hands, experiments in which the Holm procedure would have found significant results that LSD missed. But unless the background information that led to the choice of LSD is substantially changed, the researcher must be comforted by the knowledge that the gambling system that is being employed will in the long run yield errors at the low prespecified rate. On the other hand, if the researcher chooses between LSD and Holm, say, only after the data are seen, the control of error rates is lost. As Miller wrote (1994), "Of course, we can be less zealous, and criticize more mildly. That will not disqualify the proposals that would survive harsher criticism...But it will inevitably compromise the rationality of the decision-making process" (p. 43).

Other well-known examples of the price paid in violating predesignation involve the choice of a one-tailed test (and direction) after the results are known or the choice of a significant covariate for use in an analysis of covariance in the same data set, both of which would increase the Type I error rate. Freedman (1983) similarly found that screening for potential predictors in regression analysis before a final model is fit and tested results in inflated Type I error rates (this result applies to the previous example of covariate choice), and Zimmerman (1996) showed that choosing between Student's *t* test and the Welch (1947) test on the basis of a test of homogeneity of variance results in a two-stage procedure whose

Type I error rates are inflated. Similar problems would arise when the choice between analysis of covariance and analysis of variance is made on the basis of results of tests for baseline differences, (This is especially peculiar when the baseline test is performed even when *random assignment was used*, because in that case the only conclusion to draw is that the randomization was not successful. Should we redo the randomization until we like the results?) or when the choice between the *t* test and a particular form of nonparametric test is made on the basis of the skewness and kurtosis of the dependent variable in the current sample.

The reason that error rates are changed as a result of any similar two-stage procedure is that the first stage test incurs its own errors, which are then compounded in the second stage. Consider Zimmerman's results. If the population variances are equal and the other assumptions of the *t* test hold, then Student's *t* test is optimal in holding its Type I error rate and yielding desired power. But the error characteristics of the *t* test are based on all possible samplings, some of which will yield two samples with apparently different variances. If, in this case, the preliminary test commits a Type I error of its own, the Welch test used at the second stage has lower power than it should, and these cases are also removed from the sampling distribution of the *t* test. The *t* is left to operate only on samples whose variances are too close. Conversely, if the population variances are unequal, a Type II error at the first stage results in the use of the *t* test when it is inappropriate, yielding an inflation of the Type I error rate of the method.

Mayo (1993) also pointed out that Pearson, whom she said shied away from Neyman's notion of inductive behavior, 'specifically denied that the tests are to be used as automatic routines for testing claims' (p. 171). Indeed, in this regard, Neyman (1957) criticized Fisher's significance testing approach of having an automatic character in apparently always selecting a one per cent *p*-value as the cutoff for significance, concluding, "There are weighty arguments against this automatism. In fact, it appears desirable to determine the level of significance in accordance with quite a few circumstances that vary from one particular problem to the next" (p. 12). These would

include a consideration of the severity of the errors, both Type I and Type II. Rosnow and Rosenthal (1989, p. 1277) may have been right in this connection when they wrote, "Surely, God loves the .06 nearly as much as the .05", but once they have decided in advance of experiment on a value that would not be too displeasing to the statistical deity, they must ensure that the methods they choose control the error rate at this level.

Mayo (1993) observed that predesignation is only called for when violating predesignation would conflict with the goal of controlling the error probabilities. One example of the use of changing error rates mid-experiment that does not affect the overall properties of the test of a theoretical hypothesis is seen in the context of multiple comparisons. A family is defined as the set of comparisons, the significance of any one of which would lead to the conclusion that the theory has been discredited.

Any contrast whose significance does not impinge on the truth of the theory under test is not part of the family. Darlington's (1990) notion of conceptual dependence, to be distinguished from statistical dependence, among contrasts that constitute a family may be helpful in deciding whether or not contrasts belong to a family. Because methodology must be committed to controlling the rate at which the theory is falsely rejected, all legitimate multiple comparison procedures do so successfully, usually through the use of the Dunn-Bonferroni or the improved Dunn-Sidak procedure. (The Bonferroni inequality is due to Boole. Cox, 1977, suggested a sequential adjustment of alpha like the one that is due to Holm, 1979. He gave credit for the suggestion to test the most significant comparison at a Dunn-protected alpha to Tippett in 1931, whereas O'Neill and Wetherill, 1971, call the Dunn-Bonferroni procedure Fisher's Significant Difference method, attributed to Fisher, 1935. For some reason, Dunn's name is too often not included in references to these methods of error rate control.)

Control at the familywise level assures that the probability that one or more of the comparisons is falsely rejected is at most the desired alpha. Because the false rejection of one

or more of the comparisons would lead to the false discreditation of the theory under test, it is this error rate that must be controlled. Any of the sequentially rejective testing procedures, such as those of Holm (1979) or Shaffer (1986), adjusts the Type I error rate assigned to the test of particular comparisons as a function of the results that have been obtained prior to the test of the particular comparisons. This is legitimate, however, because the rate of false discreditation of the theory is still controlled at the desired level, which itself must be predesignated.

Recently, some interest has been shown in the false discovery rate (FDR) multiple comparison procedure of Benjamini and Hochberg (1995). The FDR is the expected proportion of rejections that are false. Shaffer (1995) suggested that a common misconception, that alpha refers to the proportion of the rejected hypotheses that have been falsely rejected, may have been the reason for the interest in defining and controlling FDR. Benjamini and Hochberg (1995) concluded that familywise (FWE) control is important "when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is" (Benjamini & Hochberg, 1995, p. 290), as, of course, did Peirce and Neyman and Pearson. Benjamini and Hochberg (1995) showed that when all of the hypotheses associated with the multiple comparisons are true, and so the omnibus null hypothesis is true, FDR is equal to FWE, and so in this crucial circumstance, the two procedures are equally viable.

There are other circumstances, Benjamini and Hochberg (1995) felt, in which the less stringent control of FDR is acceptable, such as in exploratory analyses, especially screening problems in which it is desired to obtain as many potential discoveries as possible, but at a controlled rate so as not overly to burden the later confirmatory stage. When considering the different approaches that may be used in exploratory as compared with confirmatory analyses, it is helpful to place the analyses in the context of Peirce's abductions and inductions or of Popper's conjectures and refutations. Because there is an open admission policy toward hypotheses, there is no need for any conjectured relationship to pass a preliminary test, except for

reasons of economy. In the abductive phase, then, any level of alpha can be used that suitably reduces the number of variables later to be tested in an independent study, even values far higher than the conventional five percent level. In the confirmatory stage, however, it is absolutely essential to decide on low and predesignated values of the Type I and Type II error rates, so that the tests are as severe as possible.

#### Satisficing

In order to test a theory in isolation, instead of as a mix of theory, initial conditions, and auxiliary theories, one must specify in advance of experiment that aspect of the theory that is under test and to assign the remainder, including theories of measurement, to unproblematic background knowledge. To deal with the theory-ladenness of observations, one must remember that the observations are interpreted in terms of theories, including the theory under test. In order to subject the theory to a severe test, we must specify in advance of the experiment what the potential falsifiers of the theory will be, what observational outcomes of the experiment will cause us to regard the theory as falsified.

One of Peirce's rules regarding induction, the inferential method by which hypotheses are tested, is that of predesignation: the property for which a sample is proposed must be specified before sampling, for otherwise "it will always be possible to find some character, however obscure, in which the instances sampled agree, and whether the same proportion of the entire class...has the property will be simply a matter of accident" (Buchler, 1939, p. 246). Indeed, without predesignation, "the induction can serve only to suggest a question, and ought not to create any belief" (Peirce, 1883, p.436).

Peirce (1958) wrote, "The essential thing is that it shall not be known beforehand, otherwise than through conviction of the truth of the hypothesis, how these experiments will turn out" (p. 58). In this regard, Berkson's (1938) observation is pertinent, that if "the result of the...test is known, it is no test at all!" (p. 537). But as discussed previously, it *is* known that the probability associated with a universal generalization is zero. Recall that in Peirce's

view, no theory is true, that Ritchie showed that the probability of any inductive generalization is zero, and that Carnap found that the degree of confirmation of all genuinely universal propositions was zero. Additionally, Peirce claimed that laws of Nature, expressed as simple formulae relating physical phenomena, "are not usually, if ever, exactly true" (Peirce, 1878, p. 334), and finally, Lakatos (1978) opined "that precise particular numerical predictions would have zero measure" (p. 139). Such views are not only expressed by philosophers, and the transfer to statisticians' views concerning the null hypothesis is fairly straightforward. For example, Kempthorne (1976) similarly offered that "A potentially mystifying aspect of this process is that no one, I think, really believes in the possibility of sharp null hypotheses—that two means are absolutely equal in noisy sciences" (p. 772), and Anscombe (1956) wrote that "no one expects any scientific theory to be complete and exact (p. 25).

There are those who defend the possibility of the truth of the point null hypothesis. For instance, Frick (1995) offered as an example of a true point null hypothesis one involved in testing for evidence of extrasensory perception (ESP), and Wainer (1999) considered the case of measuring the speed of light in two reference frames, wherein it is hypothesized that light speed is the same in both experiments. Of note is the fact that the claimed truth of both of these point null hypotheses is based on the assumption of truth of the theories under test, dubious at best given the fallible nature of all knowledge. In terms of the test involving the speed of light, it has been conjectured (Webb et. al., 2001) that certain physical constants such as the speed of light, Planck's constant, and the charge of the electron have been decreasing with time. And if the speed of light were decreasing, then the hypothesis that the two experiments would yield the same value would be false, unless the experiments were conducted simultaneously, again difficult according to the special theory of relativity. The point to be emphasized is that the falseness of point null hypotheses is consistent with the fallibility of theories.

In the case of Frick's ESP example, assume for the sake of argument that ESP is

indeed not possible. In order to test this hypothesis, a person is assigned to guess pictures drawn on a set of cards that are held up in a random order, and the actual content of the card and the guess are recorded. It would be expected that if the cards are selected and the guesses are made at random, there would be zero correlation between them. Unfortunately, neither the guesses nor the card selection are truly random. Diaconis and Mosteller (1989) pointed out that “subjects guess in a notoriously nonrandom manner” (p. 856). Similarly, the order of card selection would be made on the basis of a random device, say a pseudo-random number generator, whose properties are excellent but not perfect. Indeed, MacLaren (1992) showed that the usable length of a pseudorandom sequence was the two-thirds power of its period, after which the uniformity of the sequence no longer conforms to that of a true random sequence. Therefore, the nonrandom sequences of guesses and cards selected will evidence a nonzero correlation. In any experiment, not only must the theory under consideration be true in all respects, but all other aspects of the conditions of experiment would have to be perfectly controlled in order that the value specified in the point null hypothesis be true. This is not at all likely to occur.

This is not to say that it can not happen. The complement to Peirce’s previously cited insight that there is a difference between certainty and a probability of unity is that an event whose probability is zero is not impossible. Consider being handed a lottery ticket. If there are a finite number of possible winners, then you have a finite probability of holding the ticket with the winning number. But if the population of possible winning numbers is truly infinite, then your probability of winning is zero, despite your having an actual ticket in your hand. Analogously, although it is not impossible that the numerical value specified in a point null hypothesis is equal to the population parameter, the probability that they are equal for an infinite population is zero.

As a possible solution to the dilemma posed by false point null hypotheses, Lakatos (1978) suggested, “One could...argue...that confirmation theory should be further restricted to predictions within some finite interval of error

(p. 139). Similarly, Anscombe (1956) concluded that “we expect some discrepancy between the deduced theoretical hypothesis and our observations. We wish to know if the agreement of observation with hypothesis is *good enough*” (p. 25). This notion of specifying a range within which an effect is essentially zero corresponds to Simon’s (1957) principle of satisficing and Serlin and Lapsley’s (1985) good-enough principle. As an example of the application of the satisficing principle, consider the eclipse experiment in which Einstein’s General Theory of Relativity was found to have greater predictive power than Newton’s theory (Dyson et. al., 1920). The conclusion that light seemed to be bent by a gravitational object according to Einstein’s theory was acclaimed by Thomson (1919) as the most important result obtained in connection with the theory of gravitation since Newton’s day” (p. 389). Yet the average of the four widely differing experimental values was off by 10% from theoretical prediction. When asked about the discrepancy, Einstein said that for the expert, this thing is not particularly important.

It is felt that our best theories are close to the truth, that is, that they evidence verisimilitude, and perhaps that over time our theories become closer approximations to the truth. It is necessary to shift our focus to providing a method that allows the conclusion that the theory under test is better than the old one, or that a single prediction is closer to the truth, rather than simply that the difference is nonzero or that the prediction is in error. We could, of course, be wrong. But the emphasis here is on drawing a conclusion concerning the magnitude of an effect. As Anscombe (1956) wrote in this regard, “When testing a theoretical hypothesis, should we not in any case begin by treating the problem as one of estimation, by estimating the magnitude of departure from the theoretical hypothesis” (p. 25). Often, the hypothesis test and the estimation of magnitude are considered separate parts of the analysis. For example, Yates (1948) noted, “The first point that struck the practical man was that experiments in general performed two different functions, one being to test the significance of a certain hypothesis, and the other to estimate the magnitude of the deviation from that hypothesis

if, in fact, it was found to be, or was suspected of being, untrue” (p. 204).

One reason for this apparent disconnect between hypothesis testing and estimation by confidence interval is that the traditional point null hypothesis only allows the conclusion that the parameter is not exactly as specified, whereas the essential information to be obtained in an experiment regards whether the parameter is outside of the good-enough region. Unfortunately, the classical Neyman-Pearson confidence interval can not answer this question well. In the traditional case, it is posited that the test statistic has a certain distribution, given that the parameter is equal to a specific value, and the inversion of this distribution yields the confidence interval for the parameter, given the observed test statistic. But the results of the hypothesis test can be significant, indicating a nonzero effect, without the confidence interval indicating that the magnitude of the effect is important.

Of course, the logic underpinning the standard confidence interval is solid. We can legitimately reason that if the population mean equals a particular value, then given the data, the confidence interval can be derived using the solid statistical principles offered by Neyman and Pearson. The logic is impeccable. But because the value specified in a point null hypothesis has zero probability of being correct, Descartes might have said, “I don't doubt the validity of your inference, only the premise.”

Equally troubling is the finding by Meeks and D'Agostino (1983) that the coverage probability of the classical confidence interval is liberal if one only constructs the confidence interval after rejection of the point null hypothesis. Instead, if the confidence interval is derived from the inversion of the distribution of the test statistic that would be used to test a range null hypothesis, the interval answers the question of interest regarding whether the magnitude of the effect is large enough, there is a nonzero probability that the range specified in the null hypothesis covers the limit to the population range, and the results of the confidence interval and hypothesis test are consistent. Hodges and Lehmann (1954) and Serlin and Lapsley (1985, 1993) provided tests of range null hypotheses that allow the

conclusion that an effect is large enough. An example of the use of a range null hypothesis test to show large effects was provided by MacCallum, Browne, and Sugawara (1996) in the context of covariance structure modeling. Examples of the use of confidence intervals that provide good-enough information are given in Steiger and Fouladi (1997), Cumming and Finch (2001), Fidler and Thompson (2001), and Smithson (2001).

In addition, range null hypotheses (and confidence intervals) can be used to examine theories that predict effects of at least a certain magnitude by allowing the disconfirming conclusion that the effect is smaller than that demanded by the theory. The bioequivalence literature introduced many tests that allow the conclusion that an effect is small, as did Serlin and Lapsley (1985, 1993), Rogers, Howard, and Vessey (1993), and Seaman and Serlin (1998). Serlin (2000) showed how such a test could be used in a Monte Carlo study to establish that a statistical procedure satisfies specified criteria for robustness. As previously indicated for the general case, in using any of these procedures, the criterion for a large enough effect or an effect that is small enough to disconfirm the theory must be predesignated.

#### Implications for future research

In his book on games of chance, according to David (1962), Cardano lamented that the facts of probability that he discovered contribute to mathematical understanding but not to the gambler. It has been shown, however, that quite to the contrary, the theory of probability is essential to a rational scientific methodology in the game against nature. Point null hypotheses, like universal theories, are quite probably false, as are the assumptions underlying statistical tests. As Cox (1958) wrote, “Assumptions that we make, such as those concerning the form of the populations sampled, are always untrue” (p. 369). It is essential, then, that we be able to examine the verisimilitude of theories through the application of severe range null hypothesis tests whose assumptions are themselves subjected to serious scrutiny. The *Journal of Modern Applied Statistical Methods* is particularly well-placed to advance statistical methodology in this regard.

In order to conduct a severe test of a hypothesis, the Type I error rate of the statistical procedure must be held as close as possible to its predesignated size, and the power of the test must not fall far from its specified level, regardless of the nature of the populations sampled. To this end, robust procedures for testing range null hypotheses have to be developed and investigated. The most difficult problem to be addressed likely will involve finding a means to incorporate the hypothesized good-enough range, expressed in actual or standardized units of the raw scale, into the distribution-free procedure.

For example, in a one-sample test that a theoretical prediction is no more than 0.2 standard deviations from the true value, the satisficing range must be introduced in both the hypothesis to be tested and the sampling distribution of the test statistic. The satisficing limit of 0.2 standard deviations must be expressed in terms of the population median for the range null hypothesis addressed by the signed-rank Wilcoxon test, and the null range must also be incorporated into the sampling distribution of the signed-rank statistic. Similar accommodations must be made in a multiple-sample, multiple-predictor, and/or multiple dependent variable test in which the null range is specified in terms of a measure of association, such as R-squared, or in terms of a function of eigenvalues or the Mahalanobis distance. For instance, if the range null hypothesis is stated in terms of the squared multiple correlation coefficient between a set of predictors and a dependent variable, what are the corresponding parameters and sampling distribution of the sample statistic in a rank regression test of the appropriate range null hypothesis?

Regardless of the nature of the hypotheses and tests, the assumptions underlying the procedures must be taken into account. In the one-sample case, asymmetric pre- and post-tests with unequal variances will yield asymmetric difference scores, which would violate the assumptions underlying the matched-pair Wilcoxon test, as would having a single asymmetric dependent variable. As with the matched-pair Wilcoxon test, the properties of the adjusted Mann-Whitney test of Fligner and Policello (1981) and the modified Kruskal-

Wallis test of Rust and Fligner (1984), which accommodate unequal variances in multiple-group tests of location, are affected by asymmetry. Although much work has been done in this regard, the properties of tests of symmetry seem to depend on other properties of the distribution, such as kurtosis (Antille, Kersting, & Zucchini, 1982; Fan & Gencay, 1995; Brizzi, 2002), and so more work in this area is needed. In addition, differing variances and covariances of sets of difference scores in a repeated measures design violate the assumptions of the Friedman test and other competitors (Harwell & Serlin, 1994). The multiple group, multiple measure design would analogously require nonparametric tests of sphericity and homogeneity of covariance matrices, as would the test of identity of regression lines and the test of parallelism that is used to examine hypotheses concerning moderating variables.

Most importantly, the need for range null hypothesis tests applies both to the test of theory and to the tests of assumptions. That is, the requirement of satisficing applies at all levels of the scientific endeavor. Because theories are improbable, a good-enough region must be determined in advance of experiment, so that potential falsifiers can be specified. This, in turn, requires that a range null hypothesis be tested, in order to determine if a disconfirming outcome has occurred. And the test can only be considered severe if the error probabilities are held within an acceptable range of the predesignated levels, according to a criterion of robustness.

When examining whether or not the assumptions underlying a statistical procedure are satisfied, the hypothesis to be tested concerning the assumptions must specify that the statistical model that is conjectured to apply to the data is a good enough fit, that is, that the assumptions underlying the statistical test of a substantive theory are met well enough that the statistical test itself meets its criterion of robustness. This means that a good-enough region must be specified in a range null hypothesis of the test of the validity of the assumptions underlying the statistical test of the substantive theory, and robust tests of these range null hypotheses concerning assumptions

need to be developed. To this end, Monte Carlo studies of the robustness of procedures must provide response surfaces reflecting the Type I error rate and power as a function of the inexact agreement of model and data. Pearson and Please (1975), for example, present the Type I error rates for the one- and two-tailed, one- and two-sample  $t$  tests and tests of variances in a series of graphs for varying kurtosis at specific values of skewness. A researcher could determine limits to the skewness and kurtosis that lead to the two-sample  $t$  test, say, meeting a criterion for robustness; then these limits, in turn, would be implemented in range null hypotheses in a pilot study to determine if the skewness and kurtosis of the distribution of the population from which the proposed sample is to be drawn adequately meet the requirements for robustness of the  $t$  test.

#### Conclusion

Attempts to attain knowledge as certified true belief have failed to circumvent Hume's injunction against induction. Unfortunately, Hume also showed that the search for probable knowledge, that which Locke called opinion or belief, also depended on an inductive principle. Instead, theories must be viewed as unprovable, improbable, and undisprovable (Lakatos, 1970) because, in addition to Hume's criticism of justificationism, Peirce among others showed that the empirical basis is fallible. Importantly, though, as Whewell advocated, the method of conjectures and refutations is untouched by Hume's insights.

The implication for statistical methodology is that the requisite severity of testing is achieved through the use of robust procedures, whose assumptions have not been shown to be substantially violated, to test predesignated range null hypotheses. Nonparametric range null hypothesis tests need to be developed to examine whether or not effect sizes or measures of association, as well as distributional assumptions underlying the tests themselves, meet satisficing criteria.

#### References

- Agassi, J. (1975). Subjectivism: From infantile disease to chronic illness. *Synthese*, 30, 3-14.
- Agassi, J. (1983). Theoretical bias in evidence: A historical sketch. *Philosophica*, 31, 7-24.
- Anscombe, F. J. (1956). Discussion on Dr. David's and Dr. Johnson's paper. *Journal of the Royal Statistical Society*, Series B, 18, 24-27.
- Antille, A., Kersting, G., & Zucchini, W. (1982). Testing symmetry. *Journal of the American Statistical Association*, 77, 639-646.
- Bartley, W. W. III. (1984). *The retreat to commitment*. LaSalle, Illinois: Open Court.
- Bell E. T. (1937). *Men of mathematics*. New York: Simon & Schuster.
- Bellhouse, D. (1993). The role of roguery in the history of probability. *Statistical Science*, 8, 410-420.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, 57, 289-300.
- Bennett, J. H. (Ed.) (1990). *Statistical inference and analysis*. Oxford: Clarendon Press.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33, 526-542.
- Bradley, J. (1971). *Mach's philosophy of science*. New York: Oxford University Press.
- Brizzi, M. (2002). Testing symmetry by an easy-to-calculate statistic based on letter values. *Developments in Statistics*, 17, 63-74.
- Buchler, J. (1939). *Charles Peirce's empiricism*. New York: Harcourt, Brace and Company.
- Burt, E. A. (1924). *The metaphysical foundations of modern physical science*. London: Routledge.
- Clark, G. H. (1957). *Thales to Dewey*. Boston: Houghton-Mifflin.

- Cohen, R. S. (1970). Ernst Mach: Physics, perception and the philosophy of science. In R. S. Cohen & R. J. Seeger (Eds.), *Boston studies in the philosophy of science, Volume VI*, pp. 126-164. Dordrecht-Holland: Reidel.
- Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29, 357-372.
- Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics*, 4, 49-70.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532-74.
- Darlington, R. B. (1990). *Regression and Linear Models*. New York: McGraw Hill.
- David, F. N. (1962). *Games, Gods & Gambling*. New York: Hafner.
- Descartes, R. (1642/1927). *Meditations*. In R. M. Eaton (Ed.), *Descartes selections*. New York: Charles Scribner's Sons.
- Diaconis, P., & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, 84, 853-861.
- Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London*, 220, 291-333.
- Epstein, R. A. (1977). *The theory of gambling and statistical logic*. New York: Academic Press.
- Fan, Y., & Gencay, R. (1995). A consistent nonparametric test of symmetry in linear regression models. *Journal of the American Statistical Association*, 90, 551-557.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, 61, 575-604.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. New York: Macmillan.
- Fligner, M.A., & Policello, G. E. III (1981). Robust Rank Procedures for the Behrens-Fisher Problem. *Journal of the American Statistical Association*, 76, 162-168.
- Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37, 152-155.
- Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, 23, 132-138.
- Fuller, B. A. G. (1938). *A history of philosophy*. New York: Henry Holt and Company.
- Garber, D. (1995). Apples, oranges, and the role of Gassendi's atomism in seventeenth-century science. *Perspectives on Science*, 3, 425-428.
- Gillies, D. (1993). *Philosophy of science in the twentieth century*. Oxford: Blackwell.
- Harris, J. F. (1992). *Against relativism*. LaSalle, IL: Open Court.
- Harwell, M. R., & Serlin, R. C. (1994). A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis*, 17, 35-49.
- Hodges, J., & Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society (B)*, 16, 261-268.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.
- Jeffreys, H. (1939). *Theory of probability*. London: Oxford University Press.
- Joy, L. S. (1995). Rationality among the friends of truth: The Gassendi-Descartes controversy. *Perspectives on Science*, 3, 429-449.
- Kadane, J. B. (1976). For what use are tests of hypotheses and tests of significance, introduction. *Communications in Statistics (A)*, 5, 735-736.
- Kempthorne, O. (1976). For what use are tests of significance and tests of hypothesis. *Communications in Statistics, Part A*, 5, 763-777.
- Kiernan, J. F. (2001). Points on the path to probability. *The Mathematics Teacher*, 94, 180-183.



Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Imre Lakatos & Alan Musgrave (Eds.), *Criticism and the growth of knowledge*, 91-196. Cambridge: Cambridge University Press.

Lakatos, I. (1978). Newton's effect on scientific standards, in J. Worrall & G. Currie (Eds.): *The methodology of scientific research programmes*, 193-222. Cambridge: Cambridge University Press.

Levertoy, D. (1961). Matins. *The Jacob's ladder*. New York: New Directions.

Levi, I. (1967). *Gambling with truth*. New York: Knopf.

Levi, I. (1980). Induction as self correcting according to Peirce. In D. H. Mellor, *Science, belief, and behavior: Essays in honor of R. B. Braithwaite*, 127-140. Cambridge: Cambridge University Press.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130-149.

MacLaren, N. M. (1992). *Journal of Statistical Computation and Simulation*, 42, 47-54.

Mayo, D. G. (1993). The test of experiment: C. S. Peirce and E. S. Pearson. In E. C. Moore (Ed.), *Charles S. Peirce and the philosophy of science*, 161-174.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Medawar, P. (1974). Hypothesis and imagination. In P. A. Schilpp (Ed.), 274-291. LaSalle, IL: Open Court.

Meeks, S. L. & D'Agostino, R. B. (1983). A note on the use of confidence limits following rejection of a null hypothesis. *The American Statistician*, 37, 134-136.

Miller, D. (1980). Can science do without induction? In L. J. Cohen & M. Hesse (Eds.), *Applications of inductive logic*, 109-129.

Miller, D. (1982). Conjectural knowledge: Popper's solution of the problem of induction. In P. Levinson (Ed.), *In pursuit of truth*, 17-49. New Jersey: Humanities Press.

Miller, D. (1994). *Critical rationalism*. Chicago: Open Court.

Miller, D. (2002). Induction: A problem solved. In J. M. Böhm, H. Holweg, & C. Hoock: *Karl Popper's kritischer rationalismus heute*, 81-106. Tübingen: Mohr Siebeck.

Milnor, J. (1954). Games against nature. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes*, 49-59.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, 22, 267-305.

Musgrave, A. (1993). Popper on induction. *Philosophy of the Social Sciences*, 23, 516-527.

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *International review of statistics*, 25, 7-22.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society*, A, 231, 289-337.

O'Neill, R., & Wetherill, G. B. (1971). The present state of multiple comparison methods. *Journal of the Royal Statistical Society*, Series B, 33, 218-250.

Owen, D. (1993). Locke on reason, probable reasoning, and opinion. *The Locke Newsletter*, 24, 35-79.

Pearson, E. S., & Sekar, C. C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28, 308-320.

Pearson, E. S., & Please, N. W. (1975). Relation Between the Shape of Population Distribution and the Robustness of Four Simple Test Statistics. *Biometrika*, 62, 223-241.

Peirce, C. S. (1868). Some consequences of four incapacities claimed for man. *Journal of Speculative Philosophy*, 2, 140-157.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. In C. J. W. Kloesel (Ed.), *Writings of Charles S. Peirce*, 3, 323-338. Bloomington, Indiana: Indiana University Press.

Peirce, C. S. (1883). A theory of probable inference. In C. J. W. Kloesel (Ed.), *Writings of Charles S. Peirce*, 4, 408-450. Bloomington, Indiana: Indiana University Press.

- Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce*. Edited by A. W. Burks. Vol. VII.: *Science and Philosophy*. Cambridge: Harvard University Press.
- Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Popper, K. (1962). *Conjectures and refutations*. New York: Basic Books.
- Popper, K. (1985). *Realism and the aim of science*. From W. W. Bartley III (Ed.), *Postscript to the logic of scientific discovery*. London: Routledge.
- Quine, W. V. O. (1951). Two dogmas of empiricism. Reprinted in *From a logical point of view*. (2nd rev. ed.). Harper Torchbooks, 20-46.
- Radnitzky, G. (1982). Popper as a turning point in the philosophy of science: Beyond foundationalism and relativism. In P. Levinson (Ed.), *In pursuit of truth*, 64-80. Sussex: Harvester Press.
- Reichenbach, H. (1951). *The rise of scientific philosophy*. Berkeley: University of California Press.
- Ritchie, A. D. (1926). Induction and probability. *Mind*, 35, 301-318.
- Rogers, J., Howard, K., and Vessey, J. (1993). Using Significance tests to evaluate equivalence between experimental groups. *Psychological Bulletin*, 11, 553-565.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- Russell, B. (1945). *A history of western philosophy*. New York: Simon and Schuster.
- Rust, Steven W., & Fligner, Michael A. (1984). A modification of the Kruskal-Wallis statistic for the generalized Behrens-Fisher problem. *Communications in Statistics, Part A -- Theory and Methods*, 13, 2013-2027.
- Salmon (1966). The foundations of scientific inference, in R. G. Colodny (Ed.): *Mind and Cosmos*, 135-275. Pittsburgh: University of Pittsburgh Press.
- Seaman, MA, & Serlin, RC (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, 3, 403-411.
- Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, 5, 230-240.
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, 40, 73-83.
- Serlin, R. C. & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences*, 199-228. Hillsdale, N. J.: Erlbaum.
- Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81, 826-831.
- Shaffer, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology*, 46, 561-584.
- Simon, H. A. (1957). *Models of man, social and rational*. New York: Wiley.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, 61, 605-32.
- Steiger, H. H., & Fouladi, R. T. (1996). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger, (Eds.), *What if there were no significance tests?*, 221-257. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Stigler, S. M. (1986). *The history of statistics*. Cambridge, Mass.: Harvard University Press.
- Suppe, F. (1977). *The structure of scientific theories*. Chicago: The University of Illinois Press.
- Sylla, E. D. (1998). The emergence of mathematical probability from the perspective of the Leibniz-Jacob Bernoulli correspondence. *Perspectives on Science*, 6, 41-76.
- Todhunter, I. (1865). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*. Cambridge: Cambridge University Press.
- Thomson, J. J. (1919). Joint Eclipse Meeting of the Royal Society and the Royal Astronomical Society. *The Observatory*, 42, 389-398.

Von Mises, R. (1970). Ernst Mach and the empiricist conception of science. In R. S. Cohen & R. J. Seeger (Eds.), *Boston studies in the philosophy of science*, Volume VI, 245-270. Dordrecht-Holland: Reidel.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, 4, 212-213.

Walker, H. M. (1929). *Studies in the history of statistical method*. Baltimore: Williams and Wilkins.

Watkins, J. (1978). The Popperian approach to scientific knowledge. In G. Radnitzky & G. Andersson (Eds.), *Progress and rationality in science*, 23-43. Dordrecht, Holland: Reidel.

Webb, K., Murphy, M. T., Flambaum, V. V., Dzuba, V. A., Barrow, J. D., Churchill, C. W. Prochaska, J. X., & Wolfe, A. M. (2001). Further evidence for cosmological evolution of the fine structure constant, *Physical Review Letters*, 87, 091301-1-091301-4.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 29-35.

Wettersten, J. R. (1992). *The roots of critical rationalism*. Atlanta: Rodopi.

Wettersten, J. R. (1993). Rethinking Whewell. *Philosophy of the Social Sciences*, 23, 481-515.

Wiener, P. P. (1972). *Evolution and the founders of pragmatism*. Philadelphia: University of Pennsylvania Press, Inc.

Yates, F. (1948). Discussion on Mr. Anscombe's paper. *Journal of the Royal Statistical Society, Series A*, 111, 204-205.

Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *The Journal of General Psychology*, 123, 217-231.

## Chronic Disease Data And Analysis: Current State Of the Field



Ralph D'Agostino, Sr.  
Boston University



Lisa M. Sullivan  
Boston University

---

Chronic disease usually spans years of a person's lifetime and includes a disease free period, a preclinical, or latent period, where there are few overt signs of disease, a clinical period where the disease manifests and is eventually diagnosed, and a follow-up period where the disease might progress steadily or remain stable. It is often of interest to investigate the relationship between risk factors measured at a point in time (usually during the disease free or preclinical period), and the development of disease at some future point (e.g., 10 years later). We outline some popular designs for the identification of subjects and discuss issues in measurement of risk factors for analysis of chronic disease. We discuss some of the complexities in these analyses, including the time dependent nature of the risk factors and missing data issues. We then describe some popular statistical modeling techniques and outline the situations in which each is appropriate. We conclude with some speculation toward future development in the area of chronic disease data and analysis.

Keywords: Chronic disease, cardiovascular disease, Framingham Heart Study, logistic regression analysis, longitudinal data, missing data, mixed models, survival analysis

---

### Introduction

A chronic disease is a disease first characterized by a development period or latent period in

---

Ralph B. D'Agostino, Sr., is Professor of Mathematics/Statistics, Public Health and Law. He is a fellow of the American Statistical Association and the Cardiovascular Epidemiology section of the American Heart Association. Email: [ralph@bu.edu](mailto:ralph@bu.edu). Lisa M. Sullivan is Associate Professor of Biostatistics in the School of Public Health, Associate Professor of Mathematics and Statistics in the College of Arts and Sciences, and Associate Professor of Medicine in School of Medicine. She is Co-Director of the Graduate program in Biostatistics at Boston University. Email: [lsull@bu.edu](mailto:lsull@bu.edu).

which the disease progresses subclinically. The latent period can be extensive in time. For example, in cardiovascular disease, build up of plaque in the arteries can begin in childhood. During this latent period the person often displays no overt effects or problems. Then the disease manifests itself in a clinical phase.

With cardiovascular disease, this may begin with a myocardial infarction (heart attack) where the heart suffers permanent injury due to the blockage caused by the plaque. After the appearance of the clinical phase, the affected person (or host) may follow a course that leads to little or substantial deterioration and possibly death.

In this example of cardiovascular disease, the clinical phase is initiated by a clinical event, a heart attack, and then followed

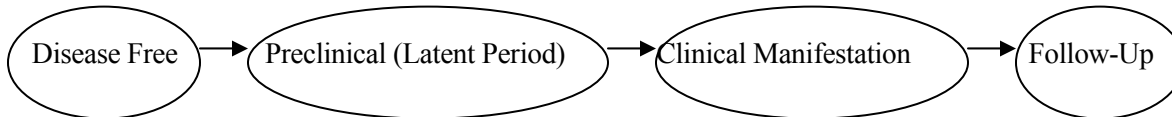
by a post event phase where there may be a general weakening of the body which increases the risk of subsequent cardiovascular events such as a second heart attack or a stroke resulting in death.

Lung cancer is an example of another chronic disease. Here the subclinical, latent period can consist of lung tumors developing over a period of more than 10 years before clinical manifestation and diagnosis. After diagnosis, there can be periods of stabilization, remission and progression. AIDS is still another example, where the subclinical stage can be characterized by a positive HIV infection. The clinical manifestation of AIDS

may then appear followed by a series of infections, increased deterioration and ultimately death. Alzheimer's disease provides an example where the distinction between the preclinical stage and clinical stage is blurred. In the preclinical phase, there is a progressive decline in cognitive function, especially noted in short term memory, and often personality changes. These ultimately lead to a stage where the person is unable to care for him or herself. The diagnosis of Alzheimer's disease often results when the person is debilitated and other forms of dementia (e.g., caused by a series of strokes) are ruled out.

A simple model for chronic disease is as follows:

(1)



Interest focuses on all four components. Each presents detailed and sophisticated modeling, data collection and analytic issues. Consider, for example, the 'Disease Free -> Preclinical (Latent Period) -> Clinical Manifestation'

component. This can be further refined to three submodels (shown below) where DF represents a completely disease-free state, PC represents preclinical signs and symptoms and C represents disease manifestation (clinical):

$$DF \longrightarrow PC \longrightarrow C \tag{2.1}$$

$$DF \begin{cases} \nearrow PC1 \\ \searrow PC2 \end{cases} \begin{matrix} \longrightarrow C \\ \longrightarrow C \end{matrix} \tag{2.2}$$

$$DF \longleftrightarrow PC \longrightarrow C \tag{2.3}$$

In (2.1), the disease free (DF) stage leads to the preclinical (PC) stage which in turn leads directly to the clinical stage (C). In such a situation knowledge of the preclinical stage could be useful in delaying or averting the clinical stage (C). Simple models of breast and colon cancer fit this situation. In (2.2), the disease free (DF) stage can lead to preclinical stages 1 or 2 (PC1 and PC2, respectively). PC1 does not progress to the clinical stage (C) while PC2 does. In this situation, identification of the

preclinical stage (PC) does not imply that the clinical stage (C) follows. Cervical cancer is an example of this situation. Lastly, (2.3) displays a situation where the preclinical stage (PC) may actually revert to the completely disease free (DF) stage or may lead to the clinical (C) stage.

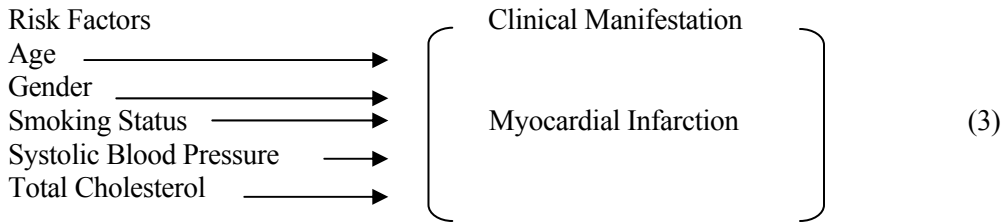
We could extend and elaborate the second component of model (1) 'Clinical Manifestation -> Follow-Up' in a similar fashion incorporating the complexities that are

involved in diagnosing the presence of the disease and the follow-up after that.

Chronic disease data and analysis questions relate to all aspects of the above (disease free, preclinical, clinical manifestation and follow-up). Good statistical approaches involve hypothesizing models for these aspects, collecting appropriate data, and then fitting and testing the appropriate models. Before fitting statistical models, biological models need to be

formulated. Both (1) and (2) above represent simple models.

One important set of models relate risk factors (RF) of a disease free individual to the probability of manifestation of the clinical stage of the disease. For example, the relationship between age, gender, smoking status, blood pressure and cholesterol to the development of a myocardial infarction could be modelled as:



To turn this into a statistical model one needs to decide how to identify appropriate (disease free) subjects, how many subjects to sample, when to measure the risk factors and how long to follow them. The latter item of follow-up is to ensure that a sufficient number develop a myocardial infarction so the components (or parameters) of the mathematical model can be estimated with good precision.

In a later part of this article we discuss in more detail the methods of statistical modeling for chronic disease. We discuss some popular designs for studies of chronic disease and we use cardiovascular disease as an example throughout the discussion. We review some of the methodologic issues that arise in studies of chronic disease and outline some popular statistical modeling and analysis techniques. We conclude with some speculation towards future developments. In the next section we present an example to motivate the discussion that follows.

2. Motivation: Cardiovascular Disease Example

Consider a study of cardiovascular disease, in particular a study of the risk factors associated with the development of cardiovascular disease. A first challenge is to understand the outcome, and in particular the conditions that should be considered part of the

outcome and how they should be measured. A second challenge is to determine which risk factors should be measured and how frequently they should be measured in the study subjects. A related challenge is the specification of the appropriate statistical model to relate candidate risk factors to the outcome. In the following we illustrate the complexities of each step using cardiovascular disease as an example.

*Defining the Outcome.* Cardiovascular disease includes a number of conditions and is a major cause of morbidity and mortality worldwide. The most common serious cardiovascular disease is coronary heart disease (also called cardiac ischemia, defined as insufficient blood supply due to atherosclerosis of the coronary arteries). It consists of myocardial infarction (heart attack), which is direct damage to the heart, coronary deaths, and angina (persistent chest pain due to cardiac ischemia). Cardiovascular disease also includes other conditions such as stroke (or brain attack), and peripheral artery disease (circulation problems often in the calves). Cardiovascular disease is believed to have a long preclinical or latent stage.

For example, patients with coronary heart disease (CHD) are diagnosed (and enter the clinical stage) in a variety of ways. One patient may present with angina at an early

stage while another may suffer a heart attack after an otherwise asymptomatic history. Accurate determination of a cardiovascular event is critical, and the technologies to determine specific events are evolving over time. At one time an MI was mainly diagnosed by electrocardiogram. Now it is standard to use enzyme tests (e.g., SGOT and CPK). Often chronic disease outcomes include condition-specific mortality (e.g., death due to cardiovascular disease). In such cases, elaborate protocols are required to ascertain outcome status. These include, in some cases, reviewing death certificates and/or hospital records. Determining cause of death can be further complicated by incomplete or ambiguous specification of the cause of death by the medical personnel evaluating the death.

*Specifying the Risk Factors and the Data Collection Schedule.* Determining the risk factors associated with the development of chronic disease (e.g., cardiovascular disease) requires an understanding of the biological complexity of the disease, some of which might change over time. Generally, studies of cardiovascular disease consider the following risk factors: gender, age, blood pressure, cholesterol, smoking status, and history of diabetes. Cardiovascular diseases span decades of individuals' lives (from the preclinical to the clinical and follow-up stages).

Studies of cardiovascular disease often take years to complete, with the duration of the study influenced by the time it takes to observe a sufficient number of outcome events. The importance and influence of risk factors may vary over time (e.g., obesity at an early age and maintained over time may be important in leading to cardiovascular disease while the most recent blood pressure may be more important than blood pressure measured decades earlier). So, often risk factors are measured at the outset, and then repeated over the follow-up period. Investigators must decide what intervals are most appropriate to obtain repeat measurements. The interval is influenced by the stability (or lack of) of the risk factors over time.

For example, total cholesterol level is a relatively stable risk factor whereas smoking status is not. The latter would need to be measured on a more frequent basis. In recent

studies of cardiovascular disease, investigators consider genetic and environmental factors, along with a broader array of clinical risk factors. In some cases, investigators have the flexibility to add new risk factors to a data collection protocol during an ongoing study. This introduces an analytic issue in that these new risk factors will not be measured on the same schedule as the core set (i.e. those measured since the outset). In cardiovascular disease, surgical procedures have also advanced rapidly in the last two decades and include introduction of artificial aortic valves, open heart surgery, angioplasty (opening blocked arteries using balloon catheters) and regulation of heart rhythms by implanted pacemakers.

In parallel, pharmacologic treatments have become increasingly effective in treating known risk factors of cardiovascular disease (e.g., hypertension, hyperlipidemia) thereby slowing the manifestation and progression of disease. It is important to measure these interventions, which generally modify the effects of the risk factors on the development of disease, along with the risk factors themselves. Designs for studies of chronic disease and methodologic issues that arise in studies of chronic disease are discussed in detail in Section 3.

*Choosing the Correct Model.* The choice of the appropriate statistical model should be based primarily on a biological model. It should also be influenced by specific aspects of the design such as whether subjects are followed for a fixed period of time and then determined to have or not have the disease at the end of the observation period or whether subjects are followed for different amounts of time and have disease status ascertained at the end of the observation period. In a study of cardiovascular disease, a subject might die during the observation period due to cancer (or some disease other than cardiovascular disease) and at the time of death be free of cardiovascular disease. The most appropriate statistical model is one that utilizes all of the information that was measured on this person rather than exclude him or her because of the complexity of the data. Popular statistical models for studies of chronic disease are discussed in detail in Section 4.

### 3. Designs, Subject Selection and Data for Studies of Chronic Disease

The data for studies of the relationship between risk factors and development and progression of chronic disease can be prospective, retrospective or cross-sectional. Prospective study designs involve identifying individuals who are free of the disease of interest and following them over time. These studies can include repeated measurements of risk factors over time and monitoring for the development and progression of disease. The schedule for following individuals and repeating measurements depends on a number of factors including the stability of the risk factors over time and the nature of the relationship between the risk factors and disease status over time. Retrospective studies (also called case control studies) usually involve identifying two groups of individuals; those with the disease of interest (often called cases) and matches who are free of the disease of interest (often called controls).

Data are collected retrospectively usually by way of individual's recollection of prior health and risk behaviors or through medical record review. These studies are not optimal. It is usually difficult to assemble representative groups of cases and controls. Often the cases represent either the sickest (e.g., subjects enrolled through an Alzheimer's clinic) or the healthiest (e.g., those who have not died) of those affected with the disease. Further, the controls often differ in many ways from the cases, confounding the comparison of cases and controls. In addition, these studies can be subject to a number of biases (for example, recall bias or inaccurate recollection of specific behaviors or measurement based on incomplete medical records).

Cross-sectional studies are conducted at a point in time and represent concurrent risk factor and disease status. In some cross-sectional studies, individuals provide historical data on risk behaviors on the basis of recollection, thereby also subjecting these studies to recall bias.

Longitudinal cohort studies are most well suited for the analysis of chronic disease. We now describe in detail the specifics of longitudinal cohort studies and outline a well

known study of cardiovascular disease, the Framingham Heart Study.

#### 3.1. Longitudinal Cohort Studies: The Framingham Heart Study

In longitudinal cohort studies, a group or cohort of individuals is assembled at the outset. The inclusion criteria often require a set of individuals to be free of the disease of interest. This is not always the case and those with prevalent disease may be enrolled at the outset. Individuals are followed prospectively in time. Serial measurements can be taken on a predetermined schedule, often at fixed time intervals (e.g., measurements every 2 years or every 5 years). Outcome or disease status is measured over time. For those individuals who develop disease, measures of the progression or severity of disease are also taken. There are several, large longitudinal cohort studies of cardiovascular disease, probably the best known study is the Framingham Heart Study, described below.

The Framingham Heart Study began in 1948 and is one of the most ambitious and daring longitudinal medical studies ever initiated. A cohort of 5,209 individuals, 2336 males and 2873 females, was enrolled from Framingham, MA. These represented a 60% sample of the town with ages from 28 to 62 years. Multiple risk factors were measured biennially, and the study continues today with surviving participants involved for over 50 years. Major cardiovascular risk factors have been measured since the outset (e.g., blood pressure, total cholesterol and smoking status) while others have been introduced as they were hypothesized to have an impact on the development of cardiovascular disease (e.g., HDL cholesterol, LDL cholesterol, homocystene and fibrinogen). Development of cardiovascular events is recorded over time including coronary heart disease (and its components; myocardial infarction, coronary death and angina), stroke, intermittent claudication (a peripheral arterial disease), congestive heart failure and cardiovascular disease death. Intense efforts continue to be utilized to gather complete information on every subject. There are some missing data due to subjects moving from the area or discontinuing



participation (which is minimal). The total loss to follow-up is less than 3 percent. The Framingham Heart Study was expanded in 1971 to include a cohort of the offspring of the original participants and their spouses. These data allow for an investigation of the evolution of new detection technologies such as echocardiogram and carotid ultrasound and the study of the effects of genetics on development of cardiovascular and other chronic diseases such as dementia.

### 3.2. Methodological Issues in Chronic Disease Studies

There are a number of major methodologic issues that arise in longitudinal studies, two are discussed here. The first issue is based on changing definitions of risk factors and outcomes over time. For example, technological advances have resulted in better diagnostic tests for determining the presence or absence of chronic disease. Studies utilizing better diagnostic tests might observe more outcome events and possible different relationships between risk factors and disease. In some chronic diseases (e.g., diabetes) medical specialists have revised the clinical criteria for diagnosing an individual (e.g., different threshold criteria on laboratory tests).

Even the definition of myocardial infarction has changed over time. In the late 1940s, its determination was based mainly on electrocardiogram. Later, enzyme tests, SGOT and CPK, became standard components of the definition of myocardial infarction starting in the mid 1950s and proceeding during the 1960s. In other areas, more sensitive assays have been developed over time for measuring risk factors (e.g., HDL and LDL cholesterol). As modifications occur during a study, analysts must take steps to make the data as comparable over time as possible. The same applies when making comparisons to external studies, these may have employed different definitions and assays.

A second methodological issue in longitudinal studies concerns missing data. Even when intensive surveillance programs are in place, such as those used in the Framingham Heart Study, there are often situations where complete data is not gathered on every subject.

In longitudinal studies of chronic disease, there are instances where data are missing because subjects fail to show up at scheduled examinations, fail to complete certain assessments even when attending the examination, or drop out during the course of the study. These circumstances produce unequal numbers of repeated measurements on different individuals. There are several approaches for performing analysis in the presence of missing data.

First, analysis can be restricted to only those individuals with complete data. This approach is not optimal in terms of efficiency and is biased in some situations. A second approach involves imputing or ascribing values for the missing values and then analyzing the revised dataset. There are sophisticated procedures and software packages available for this imputation and subsequent analysis. This analysis can be biased and can artificially improve precision. A third approach involves analyzing the incomplete dataset (i.e., without attempting to impute values for the missing data).

Statistical techniques and associated computer software (e.g., mixed models) exist that take advantage of all available data and minimize bias that are associated with analysis restricted to only individuals with complete data or analysis of imputed data. These techniques, however, require assumptions about the non-response or the missing data mechanisms. If these assumptions are incorrect, these models can also produce biased results.

The most appropriate analytic techniques in the presence of missing data are those closely tied to the underlying missing data mechanism. When the missingness does not depend on the value of the complete or missing outcome, the data are said to be missing completely at random. Data are missing completely at random if the probability of observing a missing value does not depend on current or future data. For example, if a data monitor forgets to ask a patient if he or she has persistent chest pains (angina) the missingness has nothing to do with this subject's cardiovascular health. A less strict assumption about the missing data mechanism is one in which the missingness is related only to the data observed (and not related to unmeasured or missing data).

This missing data mechanism is called missing at random and the probability of observing a missing value depends on past data but does not depend on current or future data. For example, missing at random results when missingness is related to past cardiovascular health but is independent of unavailable current or future cardiovascular health. Data that are missing completely at random or missing at random are said to be ignorable and to produce a valid analysis it is not necessary to model the missing data mechanism explicitly. Appropriate analysis that include variables related to the mechanism for missingness produce unbiased results.

The final classification of missing data mechanisms is called nonignorable missingness. If the probability of observing a missing value depends on unmeasured current and future data, the missingness is nonignorable. An example would be a subject who fails to show up for an evaluation because his/her health has started to deteriorate. The deterioration continues, and if outcomes were measured, they would reflect the decline. When missing data are nonignorable, it is critical to model the missing data mechanism explicitly in statistical models otherwise results will be biased.

Even with these classifications for missing data and the available statistical techniques and software, there is no formal means to test which mechanism is operating in a given situation. The validity of the analysis often depends heavily upon the assumptions of the technique. Therefore, analysis and interpretation of results in the presence of missing data are often open to criticism. The best recommendation for handling missing data is to avoid it wherever possible.

#### 4. Analytic Techniques for Chronic Disease Modeling

After the sample is selected and the risk factors, the outcomes and the sampling schedule determined, mathematical/statistical modeling is needed to tie these together. Several analytic techniques can be applied to investigate this relation of the risk factors to the development and progression of chronic disease. Some of these are designed specifically to relate baseline risk factors to disease development. Some are able to exploit

the time dependent nature of the risk factors and the outcome events. We now describe some popular techniques.

#### 4.1 Logistic Regression Analysis: Dichotomous Outcome

Logistic regression analysis can examine and quantify the effects of risk factors on the development of disease. The outcome of interest is dichotomous (e.g., development or non-development of chronic disease over a time period), and the independent variables or risk factors can include continuous or discrete characteristics. The logistic regression model is of the form:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

where  $Y$  is a dichotomous outcome variable (e.g., 0=no chronic disease, 1=chronic disease) and  $p=P(Y=1)$  is the probability of a subject with the disease,  $x_1, x_2, \dots, x_p$  are the risk factors, and  $\beta_1, \beta_2, \dots, \beta_p$  are the regression parameters reflecting how the risk factors affect the log of the odds of developing disease. Logistic regression analysis is a very useful technique for analyzing dichotomous outcomes and the individual is considered the unit of analysis.

Logistic regression analysis is appropriate in studies of chronic disease where originally disease free subjects are followed for a pre-specified observation period and at the end of the observation period, each subject can be classified as having developed the disease or not. In many studies of chronic disease, there are often have a number of individuals for whom we do not have data at the end of the observation period and the last time they were observed they had not yet developed disease. Logistic regression can not deal directly with these subjects. The analysts must arbitrarily drop them from analyses or assume a disease status at the end of the observation period. The techniques described in the next section can handle this and other issues that arise in longitudinal studies of chronic disease.

#### 4.2 Survival Analysis: Time to Event Data

Survival analysis includes a set of techniques that deal with time until the event of interest occurs (e.g., onset of disease). It is often the case in studies of chronic disease that there are many patients who do not develop the disease or for whom we do not know if they ever develop the disease. This happens when the disease is rare, when patients are lost to follow-up (e.g., move away but do not develop the disease), when patients die during the observation period but are free of the disease of interest at the time of death, or when they drop out of the study (e.g., due to lack of interest).

In all of these situations, we do not have the time to the development of disease. However, these individuals can contribute a substantial amount of information (up to the end of the observed time period when we know they are disease free) – information which can be utilized through survival analytic techniques. It is this aspect of the data that distinguish survival analysis techniques from other statistical techniques.

These observations in which we know the individual is disease free for some period of time, but do not know if they developed the disease in other time periods are called censored observations. There are several different types of censoring, the most common in studies of chronic disease is right censoring. Right censored observations are observations in which we do not observe the time to event because if it occurs it occurs after the last observation point.

Some survival models are based on parametric assumptions about the distribution of the survival function, while others are not (parametric and nonparametric models, respectively). A useful method to characterize survival is by the hazard function (the instantaneous rate of developing disease). There are a number of popular parametric survival models. The exponential model is perhaps the simplest, but assumes constant hazard over time and is therefore not generally applied to chronic disease data. The Weibull distribution model is a generalization of the exponential model and is popular for analyzing chronic disease risk (e.g., cancer

risk) and the hazard function is given by the following:

$$h(t) = \lambda \gamma t^{\gamma-1}$$

where  $\lambda = -\ln(p)/t$  and  $p = P(\text{disease free at time } t)$ . The hazard at time  $t$ ,  $h(t)$ , increases as  $t$  increases for  $\gamma > 1$  and decreases as  $t$  increases if  $0 < \gamma < 1$ . The exponential model is a special case of the Weibull model with  $\gamma = 1$  (constant risk with time).

Survival analysis methods can be used to assess the effects of risk factors on the development of chronic disease. There are several models that are appropriate for this purpose. A popular parametric model for analysis of chronic disease is the accelerated failure time model whose hazard function is

$$h(t) = e^{\beta'x} h_0(e^{\beta'x} t)$$

where  $t$  reflects the time until disease onset,  $h_0(t)$  is the baseline hazard at time  $t$  (i.e., the hazard if all of the risk factors were set to zero),  $\beta'x = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ ,  $x_1, x_2, \dots, x_p$  are the risk factors, and  $\beta_1, \beta_2, \dots, \beta_p$  are the regression parameters.

A popular “nonparametric” survival analysis model is the proportional hazards model (also called the Cox regression model), and it is commonly used to assess the relative impact of a set of risk factors measured at a point in time (baseline) on survival and assumes that additive differences in risk factors are related to multiplicative changes in the hazard function.

The proportional hazards model can also be used to assess the impact of time-dependent covariates (i.e., risk factors that change over time) on the hazard function and on survival. This is a particularly useful feature of the model in studies of chronic disease as individuals may undergo procedures during the observation period which alter their prognosis. For example, an individual's risk of cardiovascular disease may change after undergoing coronary artery bypass surgery. The form of the Cox model is:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$$

where  $h(t)$  is the hazard at time  $t$ ,  $h_0(t)$  is the baseline hazard at time  $t$  (i.e., the hazard if all of the risk factors were set to zero), and as above  $x_1, x_2, \dots, x_p$  are the risk factors,  $\beta_1, \beta_2, \dots, \beta_p$  are the regression parameters reflecting how the risk factors affect the hazard. The risk factors,  $x_i$  above, can be variables measured at some baseline period or variables that vary over time (called time dependent variables). The proportional hazards model is actually a semi-parametric model because the distribution of the underlying hazard is not specified.

Estimating the risk of developing chronic disease per se or assessing the effects of a set of risk factors on the development of chronic disease may be complicated by a common situation in studies of chronic disease, namely, the competing risk of other diseases or death. For example, in studying the relation of risk factors to the development of coronary heart disease the competing risk of someone developing stroke needs to be considered. Similarly, in examining the relation of cigarette smoking to lung cancer the competing risk of developing a heart attack before the lung cancer is a real possibility.

Recently, there have been major efforts to estimate the lifetime risk of developing chronic diseases such as breast cancer, coronary heart disease and Alzheimer's disease. A major methodological issue involves the handling of death which can occur before the chronic disease, such as Alzheimer's disease, develops.

#### 4.3 Longitudinal Data Analysis: Mixed Models, Generalized Linear Models and Generalized Estimating Equations

A key feature of chronic disease data is the repeated aspect of the measurements. In longitudinal studies with multiple measurements taken on a set of individuals over time, analytic techniques must take into account the correlation between measurements taken on the same individual. An added complexity is the unbalanced nature of the data due to different numbers of

measurements taken on different subjects. We now describe some popular methods for analyzing incomplete longitudinal data; mixed models and generalized estimating equations.

Mixed models procedures assume that measurements taken over time are correlated and that regression coefficients vary randomly across subjects according to a specified distribution. In these applications, some of the effects are modeled as fixed (e.g., the effects of risk factors on outcome, called within subjects effects) and some as random (between subject effects). These mixed effects models are also referred to as random coefficients models, growth curve models or hierarchical models. They can also be extended to incorporate time-dependent covariates.

In these mixed effects models a parametric structure is assumed also for the covariances of the repeated measurements. There are many distinct structures that can be assumed, including the independence structure (all observations are independent), compound symmetry (the correlation between any two observations is equal to some common value), autoregressive, and unstructured (no specification of the structure of the correlations).

Currently available statistical computing packages offer many of these structures as options in their mixed models applications. Estimates of the fixed effects and the covariances of the random effects can be estimated using maximum likelihood using Newton-Raphson techniques or the Expectation Maximization (EM) algorithm. The estimates of the covariances are biased because they do not take into account the estimation of the fixed effects and therefore it is recommended that these be estimated using restricted maximum likelihood which produces unbiased estimates. Estimates of the standard errors of effects are robust for large samples.

Mixed models are appealing models for longitudinal data as they are flexible and handle unbalanced data in a highly efficient manner. It is important to note that these models produce consistent estimates (unbiased for large samples) only when data are missing

at random or missing completely at random. These models require careful specification of the fixed and random effects and a covariance structure. When appropriate specifications are made, the final estimates of the fixed and random effects, as well as the magnitude of the variance components are statistically correct and highly informative.

A generalized linear model is a model in which a specific link function (e.g., binomial, Poisson, Gamma) is specified to relate the mean (or expected) value of the outcome to a linear function of the risk factors. This has the effect of transforming the data to a linear model, but involves correct specification of the link or distribution of the outcome variable. Parameters of the model are estimated through maximum likelihood. The appropriateness of the estimates in a generalized linear model are highly dependent on the distributional assumptions.

Generalized estimating equations (GEE) are used to analyze correlated data (e.g., data measured on the same subject over time) that could otherwise be analyzed using a generalized linear model but require fewer distributional assumptions than generalized linear models, making them more appealing. The method of estimation is an extension of least squares.

Generalized estimating equations produce consistent estimates (unbiased for large samples) and robust standard errors for large samples. Generalized estimating equations are appropriate when interest lies in "marginal" effects (i.e., effects averaged over all individuals) rather than subject-specific effects. The approach is now available in many statistical computing packages and again requires specification of a covariance structure. It is appropriate under the assumption of data missing completely at random.

#### 4.4 Tree-Based Classification Methods

Still another set of techniques for relating risk factors to development of chronic disease are tree-based classification methods. These include a number of applications which are intuitively appealing, many of which are based

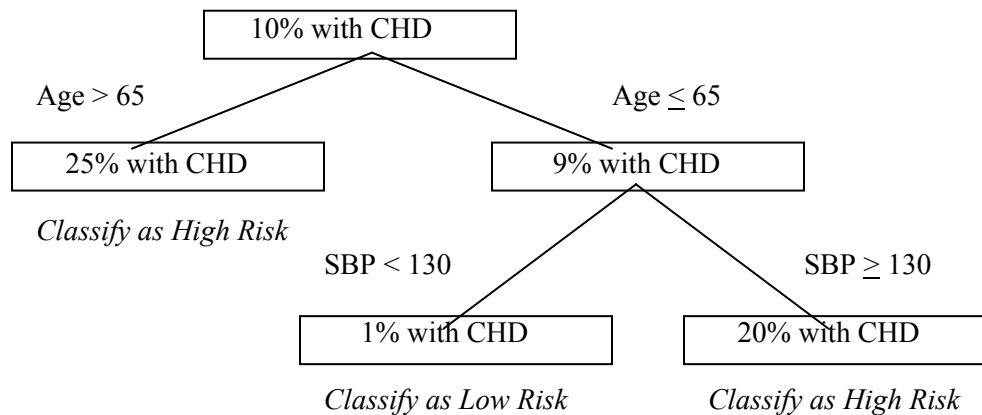
on a technique called binary recursive partitioning.

In binary recursive partitioning, a dataset is partitioned first into two distinct groups on the basis of the risk factor that best discriminates the groups in terms of disease status (present or absent). The process is recursive in that this partitioning continues until pre-specified stopping criteria are met (e.g., the final groups represent the last statistically significant splits). The outcome of these analyses is in the form of a clinical prediction rule or algorithm that resembles a tree where the branches represent splits on a risk factor.

Figure 1 illustrates a simple tree where there are two splits. The first split is on the basis of age (over 65 years versus 65 years and younger). A second split is made among those 65 years of age and younger on the basis of systolic blood pressure (less than 130 mm Hg versus 130 or more mm Hg). Persons over 65 years of age have a 25% probability of developing coronary heart disease. Persons 65 years of age and younger with systolic blood pressure less than 130 have a 1% probability of developing CHD, while persons 65 years of age and younger with systolic blood pressure of 130 or more have a 20% probability of developing CHD.

When the outcome is dichotomous (presence or absence of chronic disease) the rule can be used to classify patients, on the basis of specific criteria, as likely or unlikely to develop the disease. The criteria are based on specific values of risk factors. These models are particularly appealing to clinicians as they mirror common practice. For example, a physician might gather information from a patient on his/her risk factors (e.g., systolic blood pressure, smoking status, alcohol consumption), and may conduct a series of laboratory tests (e.g., total Cholesterol level, HDL cholesterol, triglycerides). Based on this information, the clinician can appeal to the empirical tree-based prediction rule to classify the subject as likely or not likely to develop the disease. These methods can also be used to estimate the probability that this patient will develop chronic disease.

Figure 1. Tree-Based Classification Methods: Example of A Simple Classification Tree for Coronary Heart Disease (CHD)



#### 4.5 Neural Networks

Neural network models are a large class of elaborate mathematical techniques used for developing prediction rules. They are now becoming popular methods for predicting chronic disease. They are very flexible prediction models that can accommodate large datasets (i.e., many risk factors and large sample sizes) and more complex relationships among the variables.

#### 4.6 Model Building

All of the above methods often involve a development phase and a validation phase. Investigators split a dataset into two distinct parts, one part is used for developing the model and the other part is used to evaluate how the model performs (the validation phase).

#### 5. Future Directions

The collection and analyses of chronic disease data have evolved over time to a new level of sophistication. The development of new statistical methodologies for longitudinal data analysis and analysis of complex systems, coupled with advances in statistical computing, have greatly influenced the statistical analysis of chronic disease data. As health care delivery systems continue to strive for quality, more data will be collected and available for analysis of chronic disease (and

also for acute and epidemic disease). Longitudinal data will be available on many subjects thereby allowing for more complete investigations of risk factors and interactions between risk factors.

Advances in statistical computing software will also allow for the estimation of more complex statistical models, not restricted to those which assume linear associations between risk factors and chronic disease. Finally, as more data become available on families, analysis of chronic disease will include exploration of genetic factors on the development and progression of disease.

#### References

Anderson KM, Wilson PW, Odell PM, Kannel WB (1991). An updated coronary risk profile. A statement for health professionals. *Circulation* 83, 356-362.

Beiser A, D'Agostino RB Sr., Seshadri S, Sullivan LM, Wolf PA (2000). Computing estimates of incidence, including lifetime risk: Alzheimer's disease in the Framingham Study. The Practical Incidence Estimators (PIE) macro. *Statistics in Medicine*, 19, 1495-1522.

Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, Lee PW et al. (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet* 350, 469-472.

Concato J, Feinstein AR, Holford TR (1993). The risk of determining risk with multivariable models. *Annals of Internal Medicine* 118, 201-210.

D'Agostino RB Sr., Belanger AJ, Markson EW, Kelly-Hayes M, Wolf PA (1995). Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model. *Statistics in Medicine* 14, 1757-1770.

D'Agostino RB Sr., Griffith JL, Schmid CH, Terrin N (1998). Measures for evaluating model performance. In proceedings of the Biometrics Section American Statistical Association. Biometrics Section 253-258.

Harrell FE Jr., Lee KL, Mark DB (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15, 361-387.

Hosmer DW Jr., Lemeshow S (1989). *Applied Logistic Regression*. New York: Wiley.

Knuiman MW, Vu HT, Segal MR (1997). An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *Journal of Cardiovascular Risk*, 4, 127-134.

Knuiman MW, Vu HT (1997). Prediction of coronary heart disease mortality in Busselton, Western Australia: An evaluation of the Framingham, national health epidemiologic follow-up study, and WHO ERICA risk scores. *Journal of Epidemiology and Community Health* 51, 515-519.

Laird NM. (1988). Missing Data in Longitudinal Studies. *Statistics in Medicine* 7, 305-315.

Lapuerta P, Azen PS, LaBree L (1995). Use of neural networks in predicting the risk of coronary artery disease. *Computational Biomedical Research*, 28, 38-52.

Lloyd-Jones DM, Larson MG, Beiser A, Levy D (1999). Lifetime risk of developing coronary heart disease. *The Lancet*, 353, 89-92.

Long WJ, Griffith JL, Selker HP D'Agostino RB Sr. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computational Biomedical Research*, 26, 74-97.

Segal MR, Bloch DA (1989). A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine*, 8, 539-550.

Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB Sr. (1995). A comparison of performance of mathematical predictive models for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. *Journal of Investigational Medicine*, 43, 468-476.

Seshadri S, Wolf PA, Beiser A, Au R, McNulty K, White R, D'Agostino RB Sr. (1997). Lifetime risk of dementia and Alzheimer's disease: The impact of mortality on risk estimates in the Framingham Study. *Neurology*, 49, 1498-1504.

Zeger SL, Liang KY and Albert PS (1988). Models for Longitudinal Data: A generalized Estimating Equation Approach. *Biometrics*, No. 44, 1049-1060.

Zhang H, Crowley J, Sox HC, Olshen RA (1997). Tree-Structured Statistical Methods. *Encyclopedia of Statistics*.

## Some Reflections On Significance Testing

Thomas R. Knapp  
Kailua-Kona, Hawaii



---

This essay presents a variation on a theme from my article “The use of tests of statistical significance”, which appeared in the Spring, 1999, issue of *Mid-Western Educational Researcher*.

Key words: significance tests; confidence intervals

---

### Introduction

In addition to \$.25 Senior Coffee at McDonald’s, one of the few advantages of being old at the beginning of the 21<sup>st</sup> century is that you have actually lived through certain events (World War II comes immediately to mind), rather than reading about them in history books.

An interesting statistical event that I have lived through is the controversy regarding the use of tests of significance. As David Salsburg (2001) points out in his book, *The lady tasting tea*, that controversy started in the 1930s as part of the ongoing feud between R.A. Fisher and Jerzy Neyman. It was resurrected about 35 years later with the publication of the book, *The significance test controversy*, edited by Morrison and Henkel (1970); and was revisited recently in a subsequent book entitled *What if there were no significance tests?*, edited by Harlow, Mulaik, and Steiger (1997), by a task force of the American Psychological Association (see Wilkinson, 1999), and elsewhere (e.g., Nickerson, 2000).

---

Thomas R. Knapp, Ed. D. (Harvard, 1959) is Professor Emeritus of Education and Nursing, University of Rochester and The Ohio State University. Email him at [tknapp5@juno.com](mailto:tknapp5@juno.com).

Much nonsense has been written in attempts to resolve this controversy. In what follows I would like to suggest a middle-of-the-road solution. I leave it to you, dear reader (as the late Ann Landers used to say), to decide whether or not my suggestion is more nonsense.

### Significance testing vs. hypothesis testing

Some writers (see Huberty, 1987; Huberty & Pike, 1999) distinguish between significance testing (a la Fisher) and hypothesis testing (a la Neyman & Pearson). Although the distinction is sometimes important and sometimes not, for the purposes of this paper I will not make the distinction. Here, a significance test is something one uses to test statistical hypotheses. I will also not get into null vs. nil hypotheses or one-tailed tests vs. two-tailed tests. If you are interested in such things, I recommend that you read Cohen (1965), Cohen (1994), or almost any of the late Jacob Cohen’s other work.

### Significance tests vs. confidence intervals

Since most of the controversy revolves around this matter, I will concentrate on it, along with the associated matter of “effect sizes” and what to do about them. It has often been claimed



that confidence intervals subsume significance tests: If the hypothesized value of a parameter is outside of the interval, reject it; if it is inside the interval you can't reject it. (See, for example, Steiger & Fouladi's contention that "the significance test rejects at the  $\alpha$  significance level if and only if the  $1-\alpha$  confidence interval for the mean difference excludes the value zero—1997, p. 226.) Unfortunately, it's not that simple, as Dixon and Massey (1983) and others have pointed out, especially when the parameter of interest is a population proportion or percentage, as the following example will illustrate.

#### An example

Suppose you were interested in the proportion of nurses who smoke cigarettes. (As a former holder of joint appointments in education and nursing in two different universities, I've always wondered why ANY nurses smoke!) Suppose further that you have rather limited resources and you must restrict your efforts to a relatively small population (all nurses in Rochester, New York, say) and a relatively small sample size (16, say) from same. You are familiar with some of the literature on cigarette smoking and some of the literature regarding the significance testing controversy, so you believe that you have two choices: (1) test the hypothesis that  $P$ , the population proportion, is equal to some number, say .25 (that's roughly the national average); or (2) put a confidence interval around  $p$ , the sample proportion. Let's assume that you decide on the latter choice, you draw your random sample of 16 nurses, and you find that one of the nurses smokes cigarettes.

Here is a summary of your results:

Sample  $n = 16$  Sample  $p = .0625$

Estimated standard error =

$$\sqrt{p(1-p)/n} = \sqrt{(.0625)(.9375)/16} = .0642$$

95% confidence interval =  $.0625 \pm 1.96 (.0642) = .0625 \pm .1258$ , i.e., from 0 (since you can't have a negative proportion) to .1883.

But something isn't quite right here. First of all, the normal approximation to the binomial doesn't work so well for sample sizes of 16. Secondly, the  $p$  for this particular sample is used

to estimate the population  $P$  in the calculation of the standard error, so that's a problem, since the  $P$  for this population of nurses is unknown. Finally, and perhaps most importantly, that standard error is almost certain to be an under-estimate of the "true" standard error. (It would be even worse if you just happened to draw a sample that consisted of no smokers, in which case the estimated standard error would be equal to zero!) As Wilcox (1996) and others have pointed out, you need special techniques to handle the small  $n$ , small  $p$  case.

So what? The "so what?" is that for examples like this the interval estimation approach DOES NOT subsume the hypothesis testing approach. The otherwise hypothesis-tested value of .25 is not inside the interval around your effect size of .0625 ("no effect" would be a proportion of 0), but that's not the right interval. It's too narrow. The standard error that would be used in significance testing would be a function of the .25, not the .0625.

#### Conclusion

Tom Knapp's bottom line

If you have hypotheses to test (a null hypothesis you may or may not believe a priori and/or two hypotheses pitted against one another), use a significance test to test them. If you don't, confidence intervals are fine.

I think that makes sense. Do you?

#### References

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology*, New York: McGraw-Hill.

Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, (12), 997-1003. Reprinted in L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?*. Mahwah, NJ: Erlbaum.

Dixon, W. J., & Massey, F. J. (1983). *Introduction to statistical analysis* (4<sup>th</sup> ed.). New York: McGraw-Hill.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.) (1997). *What if there were no significance tests?*. Mahwah, NJ: Erlbaum.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher*, 16 (8), 4-9.

Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. *Advances in Social Science Methodology*, 5, 1-22.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance test controversy*. Chicago: Aldine.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5 (2), 241-301.

Salsburg, D. (2001). *The lady tasting tea*. New York: Freeman

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego: Academic Press.

Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54 (8), 594-604.

## Extensions Of The Concept Of Exchangeability And Their Applications

Phillip Good  
Information Research  
Huntington Beach, California



---

Permutation tests provide exact p-values in a wide variety of practical testing situations. But permutation tests rely on the assumption of *exchangeability*, that is, under the hypothesis, the joint distribution of the observations is invariant under permutations of the subscripts. Observations are *exchangeable* if they are independent, identically distributed (i.i.d.), or if they are jointly normal with identical covariances. The range of applications of these exact, powerful, distribution-free tests can be enlarged through exchangeability-preserving transforms, asymptotic exchangeability, partial exchangeability, and weak exchangeability. Original exact tests for comparing the slopes of two regression lines and for the analysis of two-factor experimental designs are presented.

Key words: Permutation test, exchangeable, weak exchangeability, exact test, groups.

---

### Introduction

Because the permutation tests can provide exact significance levels and are powerful and distribution free, they have an enormous number of applications. See, for example, Manly(1997). The observations on which these tests are based may be drawn from finite populations or represent a particular realization of a set of random variables. Rank tests are permutation tests based on the ranks of the observations rather than their original values.

Permutation tests rely on the assumption of *exchangeability*, that is, under the hypothesis, the joint distribution of the observations is invariant

under permutations of the subscripts. Observations are *exchangeable* if they are independent, identically distributed (i.i.d.), or if they are jointly normal with identical covariances. For additional examples, see Galambos (1986) or Draper et al. (1993).

A caveat is that a set of units may be exchangeable for some purposes and not for others, depending on what is measured and the questions of interest. A simple example suggested by Draper et al (1993) is a circadian series in which observations within days are not exchangeable because of serial correlation, while observations between days (at the same point in time) are exchangeable as are the residuals from a model incorporating serial correlation.

The range of applications of these exact, powerful, distribution-free tests are enlarged below through exchangeability - preserving transforms, asymptotic exchangeability, partial exchangeability, and weak exchangeability. Original exact tests for comparing the slopes of two regression lines and for the analysis of two-factor experimental designs are presented.

### Exchangeable Variables

Let  $G\{x; y_1, y_2, \dots, y_{n-1}\}$  be a distribution function in  $x$  and symmetric in its remaining

---

Phillip I. Good is the author of five textbooks in statistics including *Permutation Tests*, *Resampling Methods*, *Applying Statistics in the Courtroom*, *Common Errors in Statistics*, and *Managers Guide to Design and Conduct of Clinical Trials*. He has published a number of short stories. See links at:

<http://users.oco.net/authors.htm> including  
<http://www.beachesbeaches.com/pinkie.html>.

arguments—that is, permuting the remaining arguments would not affect the value of G. Let the conditional distribution function of  $x_i$  given  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$  be G for all i. Then the  $\{x_i\}$  are exchangeable.

It is easy to see that a set of i.i.d. variables is exchangeable. Or that the joint distribution of a set of normally distributed random variables whose covariance matrix is such that all diagonal elements have the same value  $\sigma^2$  and all the off-diagonal elements have the same value  $\rho^2$  is invariant under permutations of the variable subscripts.

Polya's urn or contagion model variables are also exchangeable. An urn contains **b** black balls, **r** red balls, **y** yellow balls, ... and so forth. A series of balls is extracted from the urn. After the *i*th extraction, the color of the ball  $X_i$  is noted and *k* balls of the same color are added to the urn., where *k* can be any integer, positive, negative, or zero. The set of random events  $\{X_i\}$  form an exchangeable sequence. See, also, Dubins and Freedman (1979).

Transformably Exchangeable

Suggesting the concept of transformably exchangeable is the procedure for testing a non-null two-sample hypothesis  $H: F[x] = G[x-d]$ ; for if there are two sets of independent observations  $\{Z_i\}$  and  $\{Y_i\}$  with  $Z_i$  distributed as F and  $Y_i$  as G, an exact test of H can be obtained by first transforming the variables by subtracting 0 from each of the  $Z_i$ 's and *d* from each of the  $Y_i$ 's.

A set of observations (random variables) **X** will be said to be *transformably exchangeable* if there exists a transformation (measurable transformation) T, such that TX is exchangeable (Commenges, 2001).

If there are a set of observations  $\{X[t], t = 1, 2, \dots, n\}$  where  $X[t] = a + bX[t-1] + z_t$  and the  $\{z_t\}$  are i.i.d., then the variables  $\{Y[t], t = 2, \dots, n\}$  where  $Y[t] = X[t] - bX[t-1]$  are exchangeable.

Dependent non-collinear normally distributed variables with the same mean are transformably exchangeable for as the covariance matrix is non-singular, use the inverse of this matrix may be used to transform the original variables to independent (and hence exchangeable) normal ones. By applying two successive transformations, an exact permutation test can be obtained of the non-null two-sample univariate

hypothesis for dependent normally distributed variables providing the covariance matrix is known. Unfortunately, as Commenges (2001) showed, the decision to accept or reject in a specific case may depend on the transformation that was chosen.

Michael Chernick notes the preceding result applies even if the variables are collinear. Let R denote the rank of the covariance matrix in the singular case. Then, there exists a projection onto an R-dimensional subspace where R normal random variables are independent. So if there is an N dimensional ( $N > R$ ) correlated and singular multivariate normal distribution, there exists a set of R linear combinations of the original N variables so that the R linear combinations are each univariate normal and independent of one other.

Exchangeability-Preserving Transforms

Suppose it is desired to test whether two regression curves are parallel, even though the value of the intercepts are not known. Given that

$$y_{ik} = a_i + b_i x_{ik} + \varepsilon_{ik} \text{ for } i = 1, 2; k = 1, \dots, n_i$$

where the errors  $\{\varepsilon_{ij}\}$  are exchangeable. To obtain an exact permutation test for  $H: b_1 = b_2$ , the  $\{a_i\}$  are needed to be eliminated, while preserving the exchangeability of the residuals. It is known that under the null hypothesis

$$\bar{y}_i = a_i + b\bar{x}_i + \bar{\varepsilon}_i.$$

$$y' = \frac{1}{2}(\bar{y}_1 - \bar{y}_2), x' = \frac{1}{2}(\bar{x}_1 - \bar{x}_2), \varepsilon' = \frac{1}{2}(\bar{\varepsilon}_1 - \bar{\varepsilon}_2), a' = \frac{1}{2}(a_1 + a_2).$$

Define  $y'_{1k} = y_{1k} - y'$  for  $k=1$  to  $n_1$ , and  $y'_{2k} = y_{2k} + y'$  for  $k=1$  to  $n_2$ .

Define  $x'_{1k} = x_{1k} - x'$  for  $k=1$  to  $n_1$  and  $x'_{2k} = x_{2k} + x'$  for  $k=1$  to  $n_2$ .

Then

$$y'_{ik} = a' + b x'_{ik} + \varepsilon'_{ik} \text{ for } i = 1, 2; k = 1, \dots, n_i$$

Two cases arise. If the original predictors were the same for both sets of observations, that is, if  $x_{1k} = x_{2k}$  for all  $k$ , then the errors  $\{\varepsilon'_{ik}\}$  are exchangeable and the method of matched pairs can be applied; see, for example, Good (2000, p51). Otherwise, proceed as follows: First, estimate the two parameters  $a'$  and  $b$  by least-squares means. Use these estimates to derive the transformed observations  $\{y'_{ik}\}$ . Then test the hypothesis that  $b_1 = b_2$  using a two-sample comparison. If the original errors were exchangeable, then the errors  $\{\varepsilon'_{ik}\}$  though not independent are exchangeable also and this test is exact.

Now suppose

$$y_{ik} = A_i Z_k + b_i x_{ik} + \varepsilon_{ik} \text{ for } i = 1, 2; k = 1, \dots, n_i$$

where  $Z_k$  is a column vector of covariates with  $A_i$  a row vector of the corresponding coefficients. Defining  $A'_i$  as the mean of  $A_1$  and  $A_2$ , then

$$y'_{ik} = A' Z_k + b x'_{ik} + \varepsilon'_{ik} \text{ for } i = 1, 2; k = 1, \dots, n_i$$

which are analogous results for the general case.

Dean and Verducci (1990) characterized the linear transformations that preserve exchangeability. Commenges (2001) characterized the linear transformations that also preserve the permutation distribution. Clearly any transformation which preserves the ordering of the order statistics preserves exchangeability.

#### Asymptotic Exchangeability

Illustrating the concept of asymptotic exchangeability are the residuals in a two-way complete balanced experimental design. Our model is that

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where

$$\sum \alpha_i = \sum \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

and the  $\{\varepsilon_{ijk}\}$  are exchangeable. Eliminating the main effects in the traditional manner, that is, setting

$$X'_{ijk} = X_{ijk} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...},$$

the test statistic obtained is

$$I = \sum_i \sum_j (\sum_k X'_{ijk})^2,$$

which was first derived by Still and White (1981). A permutation test based on this statistic will not be exact for finite samples as the residuals

$$\varepsilon'_{ijk} = \varepsilon_{ijk} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j.} + \bar{\varepsilon}_{...}$$

are weakly correlated, the correlation depending on the subscripts. It is easy to show the Studentized correlations converge to a common value as the sample size increases, thus the residuals are asymptotically exchangeable, and the

permutation test of the hypothesis  $\gamma_{ij} = 0$  for all  $i$  and  $j$  based on  $I$  is asymptotically exact.

Romano (1990) proved asymptotic exchangeability for the two-sample comparison of independent observations with not necessarily identical distributions providing the underlying variables have the same mean and variance under the hypothesis. Baker (1995) used simulations to demonstrate the asymptotic exchangeability of the deviates about the sample median that are used in Good's test for equal variances.

#### Exchangeability and Invariance

The requirement for exchangeability in testing arises in either of two ways:

- Sufficiency—the order statistics are sufficient for a wide variety of problems.
- Invariance—the joint distribution of the observations is invariant under permutation of the subscripts.

For many testing problems, the underlying model must remain invariant under permutations of the subscripts. This can only be accomplished in many cases if the set of permutations are restricted. Recall that in the classic definition (de Finetti, 1930; Galambos, 1986) a set of  $n$  random variables is said to be *exchangeable* if the joint distribution of the variables is invariant with respect to the group  $S_n$  of all possible permutations of the subscripts.

Define the *weak exchangeability* of a set of random variables as the invariance of their joint distribution with respect to a subset of permutations. Clearly, a set of variables that is exchangeable is also weakly exchangeable.

*Exchangeability* is a necessary and sufficient condition for exactness in the classic testing problems to which permutation methods have been applied such as the 2- and k-sample tests. But in the two-factor experimental design considered in the previous section, only the error terms  $\{\varepsilon_{ijk}\}$  are exchangeable; the  $\{X_{ijk}\}$  are not.

Nonetheless, because the  $\{X_{ijk}\}$  are weakly exchangeable under any of the three null hypotheses ( $H_1: \alpha_i = 0$  for all  $i$ ,  $H_2: \beta_j = 0$  for all  $j$ , and  $H_3: \gamma_{ij} = 0$  for all  $i$  and  $j$ ), Pesarin (2001) and Salmaso (2001) were able to derive independent exact tests for each of the main effects and the interactions.

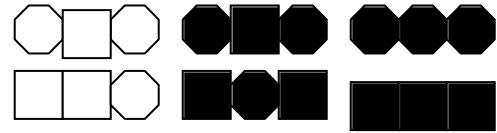
To see this, consider that the set of observations  $\{X_{ijk}\}$  may be thought of in terms of a rectangular lattice  $L$  with  $K$  colored, shaped balls at each vertex. All the balls in the same column have the same color initially, a color which is distinct from the color of the balls in any other column. All the balls in the same row have the same pattern initially, a shape which is distinct from the shape of the balls in any other row.



A 2x3 design with three observations per cell.

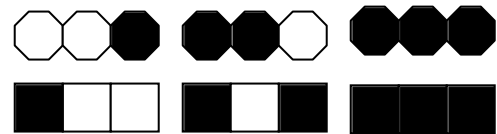
Let  $P$  denote the set of transformations that preserve the number of balls at each row and column of the lattice.  $P$  is a group.

Let  $P_R$  denote the set of exchanges of balls among rows which a) preserve the number of balls at each row and column of the lattice, and b) result in the numbers of each shape within each row being the same in each column.  $P_R$  is the basis of a subgroup of  $P$ .



A 2x3 design with three observations per cell after  $\pi \in P_R$ .

Let  $P_C$  denote the set of exchanges of balls among columns which a) preserve the number of balls at each row and column of the lattice, and b) result in the numbers of each color within each column being the same in each row.  $P_C$  is the basis of a subgroup of  $P$ .



A 2x3 design with three observations per cell after  $\pi \in P_C$ .

Let  $P_{RC}$  denote the set of exchanges of balls which preserve the number of balls at each row and column of the lattice, and result in a) an exchange of balls between both rows and columns (or no exchange at all), b) the numbers of each color within each column being the same in each row, c) the numbers of each shape within each row being the same in each column.  $P_{RC}$  is the basis of a subgroup of  $P$ . Moreover,  $P_{RC} \cap P_R = P_{RC} \cap P_C = P_R \cap P_C = I$  and  $P$  is the group generated by the union of  $P_R$ ,  $P_C$  and  $P_{RC}$ .

Define  $p[\Delta; X] = \Pi_i \Pi_j \Pi_{\kappa} f[x - \Delta_{ij}]$  where

$$\Delta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

$$\sum \alpha_i = \sum \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

and  $f$  is a density function that is continuous a.e.

Without loss of generality, it may be assumed  $\mu=0$ , or, equivalently, the set of observations  $\{X'_{ijk}\}$  obtained by subtracting  $\mu$  from each element of  $\{X_{ijk}\}$  may be used.

Suppose, now, the hypothesis  $H_1: \alpha_i = 0$  for all  $i$  holds. Then the joint distribution of the vector

$(x_{i1k}, x_{i2k}, \dots, x_{ijk})$  obtained by taking an arbitrary element from each column of the  $i$ th row is identical with the joint distribution of

$$(z - \beta_1 - \gamma_{i1}, z - \beta_2 - \gamma_{i2}, \dots, z - \beta_J - \gamma_{iJ})$$

where  $f$  is the probability density of  $z$ . The probability density of the sum of these latter elements is identical with the probability density of  $nz - \sum_{j=1}^J \beta_j - \sum_{j=1}^J \gamma_{ij} = nz$ ; that is,  $f(z/n)$ .

Under  $H_1$

- $f$  is the probability density of the mean of each of the rows of  $X$ .
- Applying any of the elements of  $\mathbf{P}_R$  leaves this density unchanged.
- Applying any of the elements of  $\mathbf{P}_R$  leaves the density of the test statistic  $F_2 = \sum_i (\sum_j \sum_k x_{ijk})^2$  unchanged.

Similarly, to test  $H_2$ , the permutation distribution over  $\mathbf{P}_C$  of any of the statistics  $F_2 = \sum_j (\sum_i \sum_k x_{ijk})^2$ ,  $F_1 = \sum_j |\sum_i \sum_k x_{ijk}|$ , or  $R_2 = \sum_j g[j] \sum_i \sum_k x_{ijk}$ , where  $g[j]$  is a monotone function of  $j$  may be used.

If  $q \in \mathbf{P}_R$  and  $s \in \mathbf{P}_C$ , then under  $H_3$ , the density of  $S_{ij} = \sum_k x_{ijk}$  is invariant with respect to  $p = qt \in \mathbf{P}_{RC}$ , and, by induction, applying any of the elements of  $\mathbf{P}_{RC}$  leaves the density of the test statistic  $S = \sum_i \sum_j (S_{ij})^2$  unchanged. As only the identity  $I$  is common to the corresponding permutation groups, the permutation tests of the three hypotheses are independent of one another.

### Partial and Weak Exchangeability

Consider a sequence of discrete random variables that represent the outcomes of a finite Markov Chain whose transition matrix is such that  $p_{ij} = p_{ji}$  for all  $i$  and  $j$ . Such a sequence is said to be *partially exchangeable* (see, for example, Zaman, 1984). If the transition matrix is connected then the sequence is also weakly exchangeable.

### References

Baker, R. D. (1995). Two permutation tests of equality of variance. *Statistics of Computation*, 5, 289-296.

Commenges, D. (2001, in press). Transformations which preserve exchangeability and applications to permutation tests. *Nonparametric statistics*.

Dean, A. M. & Verducci, J. S. (1990). Linear transformations that preserve majorization, Schur concavity, and exchangeability. *Linear algebra and its applications*, 127, 121-138.

Dean, A. M. & Wolfe, D. A. (1990). A note on exchangeability of random variables. *Statistica Neerlandica*, 44, 23-27.

Draper, D., Hodges, J. S., Mallows, C.L., & Pregibon, D. (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, A*, 156, 9-28.

Dubins, L. & Freedman, D. (1979). Exchangeable processes need not be mixtures of independent identically distributed random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 48, 115-132.

Galambos, J. (1986). Exchangeability. In S. Kotz and N. L. Johnson (Eds). *Encyclopedia of statistical sciences*, 7, 573-577. NY: Wiley.

Good, P. I. (2000) *Permutation tests*. (2<sup>nd</sup> ed.). NY: Springer.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. NY: John Wiley & Sons.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology*. (2nd ed.). London: Chapman and Hall.

Pesarin, F. (2001). *Multivariate permutation tests*. NY: Wiley.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85, 686-692.

Salmaso, L. (2003, in press). Synchronized permutation tests in 2k factorial designs. *Communications in Statistics - Theory and Methods*.

Still, A. W. & White, A. P. (1981). The approximate randomization test as an alternative to the F-test in the analysis of variance. *British Journal of Mathematical and Statistical Psychology*, 3, 243-252.

Zaman, A. (1984). Urn models for Markov exchangeability. *Annals of Probability*, 12, 223-229.

## REGULAR ARTICLES

# Twenty Nonparametric Statistics And Their Large Sample Approximations

Gail Fahoome

Educational Evaluation and Research  
Wayne State University

---

Nonparametric procedures are often more powerful than classical tests for real world data which are rarely normally distributed. However, there are difficulties in using these tests. Computational formulas are scattered throughout the literature, and there is a lack of availability of tables and critical values. The computational formulas for twenty commonly employed nonparametric tests that have large-sample approximations for the critical value are brought together. Because there is no generally agreed upon lower limit for the sample size, Monte Carlo methods were used to determine the smallest sample size that can be used with the respective large-sample approximation. The statistics reviewed include single-population tests, comparisons of two populations, comparisons of several populations, and tests of association.

Key words: nonparametric statistics, Monte Carlo methods, sample size, large sample approximation

---

### Introduction

Classical parametric tests, such as the  $F$  and  $t$ , were developed in the early part of the twentieth century. These statistics require the assumption of population normality. Bradley (1968) wrote, "To the layman unable to follow the derivation but ambitious enough to read the words, it sounded as if the mathematician had esoteric *mathematical* reasons for believing in at least quasi-universal quasi-normality" (p. 8). "Indeed, in some quarters the normal distribution seems to have been regarded as embodying metaphysical and awe-inspiring properties suggestive of Divine Intervention" (p. 5).

When Micceri (1989) investigated 440 large-sample education and psychology data sets he concluded, "No distributions among those investigated passed all tests of normality, and very

few seem to be even reasonably close approximations to the Gaussian" (p. 161). This is of practical importance because even though the well known Student's  $t$  test is preferable to nonparametric competitors when the normality assumption has been met, Blair and Higgins (1980) noted:

Generally unrecognized, or at least not made apparent to the reader, is the fact that the  $t$  test's claim to power superiority rests on certain optimal power properties that are obtained under normal theory. Thus, when the shape of the sampled population(s) is unspecified, there are no mathematical or statistical imperatives to ensure the power superiority of this statistic. (p. 311)

Blair and Higgins (1980) demonstrated the power superiority of the nonparametric Wilcoxon Rank Sum test over the  $t$  test for a variety of nonnormal theoretical distributions. In a Monte Carlo study of Micceri's real world data sets, Sawilowsky and Blair (1992) concluded that although the  $t$  test is generally robust with respect to Type I errors under conditions of equal sample size, fairly large samples, and two-tailed tests, it is not powerful for skewed distributions. Under those conditions, the Wilcoxon Rank Sum test can be three to four times more powerful. See Bridge and

---

Gail Fahoome is a Lecturer in the College of Education. Contact her at 335 College of Education, Wayne State University, Detroit, MI 48202 for all communications regarding this paper. E-mail her at [gfafoome@wayne.edu](mailto:gfafoome@wayne.edu). Her areas of expertise are Monte Carlo methods with Fortran, nonparametric statistics, and structural equation modeling.



Sawilowsky (1999) and Nanna and Sawilowsky (1998) for other examples.

The prevalence of nonnormally distributed data sets in applied studies in education and related fields has its initial impact on parametric procedures with regard to Type I errors. Thus, the immediate advantage of nonparametric procedures, such as the Wilcoxon test, is that their Type I error properties are not dependent on the assumption of population normality.

A difficulty in using nonparametric tests is the availability of computational formulas and tables of critical values. For example, Siegel and Castellan (1988) noted, "Valuable as these sources are, they have typically either been highly selective in the techniques presented or have not included the tables of significance" (p. xvi). This continues to be a problem as evidenced by a survey of 20 in-print general college statistics textbooks, including seven general textbooks, eight for the social and behavioral sciences, four for business, and one for engineering. Formulas were given for only eight nonparametric statistics, and tables of critical values were given for only the following six: (a) Kolmogorov-Smirnov test, (b) Sign test, (c) Wilcoxon Signed Rank test, (d) Wilcoxon (Mann-Whitney) test, (e) Spearman's rank correlation coefficient, and (f) Kendall's rank correlation coefficient.

This situation is somewhat improved for nonparametric statistics textbooks. Eighteen nonparametric textbooks published since 1956 were also reviewed. Table 1 contains the statistical content of the eighteen textbooks. The most comprehensive texts in terms of coverage were Neave and Worthington (1988), which is currently out of print, and Deshpande Gore, and Shanubhogue (1995).

Many nonparametric tests have large sample approximations that can be used as an alternative to tabulated critical values. These approximations are useful substitutes if the sample size is sufficiently large, and hence, obviate the need for locating tables of critical values. However, there is no generally agreed upon definition of what constitutes a *large* sample size. Consider the Sign test and the Wilcoxon tests as examples. Regarding the Sign test, Hájek (1969) wrote, "The normal approximation is good for  $N \geq 12$ " (p. 108).

Table 1. Survey of 18 Nonparametric Books

Statistic	Number of Books That Included Tables of Critical Values
<u>Single Population Tests</u>	
Kolmogorov-Smirnov Test	11
Sign Test	4
Wilcoxon Signed Rank Test	14
<u>Comparison of Two Populations</u>	
Kolmogorov-Smirnov 2-sample Test	11
Rosenbaum's Test	1
Wilcoxon (Mann-Whitney)	14
Mood Test	1
Savage Test	1
Ansari-Bradley Test	1
<u>Comparison of Several Populations</u>	
Kruskal-Wallis Test	10
Friedman's Test	9
Terpstra-Jonckheere Test	5
Page's Test	4
Match Test for Ordered Alternatives	1
<u>Tests of Association</u>	
Spearman's Rank Correlation Coefficient	12
Kendall's Rank Correlation Coefficient	10

Gibbons (1971) agreed, "Therefore, for moderate and large values of  $N$  (say at least 12) it is satisfactory to use the normal approximation to the binomial to determine the rejection region" (p. 102). Sprent (1989) and Deshpande, Gore, and Shanubhogue (1995), however, recommended  $N$  greater than 20. Siegel and Castellan (1988) suggested  $N \geq 35$ , but Neave and Worthington (1988) proposed  $N > 50$ .

The literature regarding the Wilcoxon Rank Sum test is similarly disparate. Deshpande, Gore, and Shanubhogue (1995) stated that the combined sample size should be at least 20 to use a large sample approximation of the critical value. Conover (1971) and Sprent (1989) recommended that one or both samples must exceed 20. Gibbons (1971) placed the lower limit at twelve per sample. For the Wilcoxon Signed Rank test, Deshpande, Gore, and Shanubhogue (1995) said that the approximation can be used when  $N$  is greater than 10. Gibbons (1971) recommended it when  $N$  is greater than 12, and Sprent (1989) required  $N$  to be

greater than 20. The general lack of agreement may indicate that these recommendations are based on personal experience, the sample sizes commonly accommodated in tables, the author's definition of acceptable or large, or some other unstated criterion.

There are two alternatives to tables and approximations. The first is to use exact permutation methods. There is software available that will generate exact p-values for *small* data sets and Monte Carlo estimates for *larger* problems. See Ludbrook and Dudley (1998) for a brief review of the capabilities of currently available software packages for permutation tests. However, these software solutions are expensive, have different limitations in coverage of procedures, and may require considerable computing time even with fast personal computers (see, e.g., Musial, 1999; Posch & Sawilowsky, 1997). In any case, a desirable feature of nonparametric statistics is that they are easy to compute without statistical software and computers, which makes their use in the classroom or work in the field attractive.

A second alternative is the use of the rank transformation (RT) procedure developed by Conover and Iman (1981). They proposed the use of this procedure as a bridge between parametric and nonparametric techniques. The RT is carried out as follows: rank the original scores, perform the classical test on the ranks, and refer to the standard table of critical values. In some cases, this procedure results in a well-known test. For example, conducting the *t* test on the ranks of original scores in a two independent samples layout is equivalent to the Wilcoxon Rank Sum test. (However, see the caution noted by Sawilowsky & Brown, 1991). In other cases, such as factorial analysis of variance (ANOVA) layouts, a new statistic emerges.

The early exuberance with this procedure was related to its simplicity and promise of increased statistical power when data sets displayed nonnormality. Iman and Conover noted the success of the RT in the two independent samples case and the one-way ANOVA layout. Nanna (1997, 2001) showed that the RT is robust and powerful as an alternative to the independent samples multivariate Hotelling's  $T^2$ .

However, Blair and Higgins (1985) demonstrated that the RT suffers power losses in the dependent samples *t* test layout as the

correlation between the pretest and posttest increases. Bradstreet (1997) found the RT to perform poorly for the two samples Behrens-Fisher problem. Sawilowsky (1985), Sawilowsky, Blair, and Higgins (1989), Blair, Sawilowsky, and Higgins (1987), and Kelley and Sawilowsky (1997) showed the RT has severely inflated Type I errors and a lack of power in testing interactions in factorial ANOVA layouts. Harwell and Serlin (1997) found the RT to have inflated Type I errors in the test of  $\beta = 0$  in linear regression. In the context of analysis of covariance, Headrick and Sawilowsky (1999, 2000) found the RT's Type I error rate inflates quicker than the general ANOVA case, and it demonstrated more severely depressed power properties. Recent results by Headrick (personal communications) show the RT to have poor control of Type I errors in the ordinary least squares multiple regression layout. Sawilowsky (1989) stated that the RT as a bridge has fallen down, and cannot be used to unify parametric and nonparametric methodology or as a method to avoid finding formulas and critical values for nonparametric tests.

#### Purpose Of The Study

As noted above, the computational formulas for many nonparametric tests are scattered throughout the literature, and tables of critical values are scarcer. Large sample approximation formulas are also scattered and appear in different forms. Most important, the advice on how large a sample must be to use the approximations is conflicting. The purpose of this study is to ameliorate these five problems.

Ascertaining the smallest sample size that can be used with a large sample approximation for the various statistics would enable researchers who do not have access to the necessary tables of critical values or statistical software to employ these tests. The first portion of this paper uses Monte Carlo methods to determine the smallest sample size that can be used with the large sample approximation while still preserving nominal alpha. The second portion of this paper provides a comprehensive review of computational formulas with worked examples for twenty nonparametric statistics. They were chosen because they are commonly employed and because large sample approximation formulas have been developed for them.

### Methodology

Each of the twenty statistics was tested with normal data and Micceri's (1989; see also Sawilowsky, Blair, & Micceri, 1990) real world data sets. The real data sets represent smooth symmetric, extreme asymmetric, and multi-modal lumpy distributions. Monte Carlo methods were used in order to determine the smallest samples that can be used with large-sample approximations.

A program was written in Fortran 90 (Lahey, 1998) for each statistic. The program sampled with replacement from each of the four data sets for  $n = 2, 3, \dots, N$ ;  $(n_1, n_2) = (2, 2), (3, 3), \dots, (N_1, N_2)$ , and so forth as the number of groups increased. The statistic was calculated and evaluated using the tabled values when available, and the approximation of the critical value or the transformed obtained value, as appropriate. The number of rejections was counted and the Type I error rate was computed. Nominal  $\alpha$  was set at .05 and .01. Bradley's (1978) conservative estimates of  $.045 < \text{Type I error rate} < .055$  and  $.009 < \text{Type I error rate} < .011$  were used, respectively, as measures of robustness. The sample sizes were increased until the Type I error rates converged within these acceptable regions.

### Limitations

In many cases there are different formulas for the large sample approximation of a statistic. Two criteria were used in choosing which formula to include: (1) consensus of authors, and (2) ease of use in computing and programming. All statistics were examined in the context of balanced layouts only.

Some statistics have different large sample approximations based on the presence of ties among the data. Ties were corrected using average ranks for rank-based tests, obviating tie correction formulae. For nonrank-based tests, simple deletion of ties results in a failure to adjust for variance. (A well-known example is the necessity of using a winsorized standard deviation – or some other modification to the estimate of population variance – in constructing a confidence interval for the trimmed mean when tied scores are deleted.) Nevertheless, many authors (e. g., Gibbons, 1976) indicated that adjustment for ties makes little difference for rank- or nonrank-based tests unless

there is an extreme number of ties. The issue of correcting for ties is discussed in the section below.

### Data Sets For Worked Examples In This Article

The worked examples in this study use the five data sets in Table 3 (Appendix). Some statistics converged at relatively large sample sizes. In choosing the sample size for the worked example, a compromise was made based on the amount of computation required for large samples and an unrepresentatively small but convenient sample size for presentation in this article. Therefore, a sample size of  $n = 15$  or  $N = 15$ , as appropriate, was selected, recognizing that some statistics' large sample approximations do not converge within Bradley's (1968) limits for this sample size. The data sets were randomly selected from Micceri's (1989) multimodal lumpy data set (Table 4, Appendix). Because the samples came from the same population, the worked examples all conclude that the null hypothesis cannot be rejected.

### Statistics Examined

The twenty statistics included in this article represent four layouts: (1) single population tests, (2) comparison of two populations, (3) comparison of several populations, and (4) tests of association. Single-populations tests included: (a) a goodness-of-fit test, (b) tests for location, and (c) an estimator of the median. Comparisons of two populations included: (a) tests for general differences, (b) two-sample location problems, and (c) two-sample scale problems. Comparisons of several populations included: (a) ordered alternative hypotheses, and (b) tests of homogeneity against omnibus alternatives. Tests of association focused on rank correlation coefficients.

### Results

Table 2 shows the minimum sample sizes necessary to use the large sample approximation of the critical value or obtained statistic for the tests studied. The recommendations are based on results that converged when underlying assumptions are reasonably met. The minimum sample sizes are conservative, representing the largest minimum for each test. If the test has three

or more samples, the largest group minimum is chosen. Consequently the large-sample approximations will work in some instances for smaller sample sizes. This is the smallest size per sample when the test involves more than one sample.

Table 2. Minimum Sample Size for Large-Sample Approximations.

Test	$\alpha = .05$	$\alpha = .01$
<u>Single Population Tests</u>		
Kolmogorov-Smirnov Goodness-of-Fit Test	$25 \leq n \leq 40$	$28 \leq n \leq 50$
Sign Test	$n > 150$	$n > 150$
Signed Rank Test	10	22
Estimator of Median for a Continuous Distribution	$n > 150$	$n > 150$
<u>Comparison of Two Populations</u>		
Kolmogorov-Smirnov Test	$n > 150$	$n > 150$
Rosenbaum's Test	16	20
Tukey's Test	$10 \leq n \leq 18$	21
Rank-Sum Test	15	29
Hodges-Lehmann Estimator	15	20
Siegel-Tukey Test	25	38
Mood Test	5	23
Savage Test	11	31
Ansari-Bradley Test	16	29
<u>Comparison of Several Populations</u>		
Kruskal-Wallis Test	11	22
Friedman's Test	13	23
Terpstra-Jonckheere Test	4	8
The Match Test ( $k > 3$ )	86	27
Page's Test $k > 4$	11	18
<u>Tests of Association</u>		
Spearman's Rho	12	40
Kendall's Tau	$14 \leq n \leq 24$	$15 \leq n \leq 35$

Some notes and cautionary statements are in order with regard to the entries in Table 2. The parameters for the Monte Carlo study were limited to  $n$  (or  $N$ ) = 1, 2, ... 150. The Kolmogorov-Smirnov goodness-of-fit test was conservative below the minimum value stated and liberal above the maximum value stated. Results for the Sign test indicated convergence for some distributions may occur close to  $N = 150$ . The results for the confidence interval for the Estimator of the

Median suggest convergence may occur close to  $N = 150$  only for normally distributed data. However, for the nonnormal data sets the Type I error rates were quite conservative (e.g., for  $\alpha = .05$  the Type I error rate was only 0.01146 and for  $\alpha = .01$  it was only 0.00291 for  $N = 150$  and the extreme asymmetric data set).

The Kolmogorov-Smirnov two samples test was erratic, with no indication convergence would be close to 150. Results for Tukey's test were conservative for  $\alpha = .05$  when the cutoff for the p-value was .05, and fell within acceptable limits for some sample sizes when .055 was used as a cutoff. The Hodges-Lehmann estimator only converged for normal data. For nonnormal data the large sample approximation was extremely conservative with  $n = 10$  (e.g., for the extreme asymmetric data set the Type I error rate was only 0.0211 and 0.0028 for the .05 and .01 alpha levels, respectively) and increased in conservativeness (i.e., the Type I error rate converged to 0.0) as  $n$  increased. The Match test only converged for normally distributed data, and it was the only test where the sample size required for  $\alpha = .01$  was smaller than for  $\alpha = .05$ .

These results relate to the large sample approximation of the critical values associated with those tests. These procedures work quite well with small sample sizes when tabled critical values are used. The difficulty, as noted above, is that tabled critical values are generally not available, or the implementation of exact procedures is still by far too time-consuming or memory intensive to compute with statistical software. For example, Bergmann, Ludbrook, and Spooren (2000), noted "What should be regarded as a large sample is quite vague ...most investigators are accustomed to using an asymptotic approximation when group sizes exceed 10" (p. 73). If they are correct with their perception of common practices using as few as  $n = 11$ , the results in Table 2 demonstrate that the large sample approximation of the critical value prevents the statistic from converging with nominal alpha for seventeen of the twenty procedures for  $\alpha = 0.05$ , and for nineteen of twenty procedures for  $\alpha = 0.01$ .

The vagueness of what constitutes a large sample for the purposes of using the approximation to the critical values vanishes in view of the results in Table 2. For example, with  $\alpha$

= 0.05, large for the Match test is greater than 85. This does not mean the test performs poorly and should be removed from the data analyst's repertoire if one has a smaller sample size; rather, it means the researcher is advised to have at least 86 per group before relying on the large sample approximation of the critical values.

Statistics, Worked Examples, Large Scale Approximations

Single Population Tests

Goodness-of-fit statistics are single-population tests of how well observed data fit expected probabilities or a theoretical probability density function. They are frequently used as a preliminary test of the distribution assumption of parametric tests. The Kolmogorov-Smirnov goodness-of-fit test was studied.

Tests for location are used to make inferences about the location of a population. The measure of location is usually the median. If the median is not known but there is reason to believe that its value is  $M_0$ , then the null hypothesis is  $H_0 : M = M_0$ . The tests for location studied were the Sign test, Wilcoxon's Signed Rank test, and the Estimator of the Median for a continuous distribution.

Kolmogorov-Smirnov Goodness-of-Fit Test

The Kolmogorov-Smirnov (K-S) statistic was devised by Kolmogorov in 1933 and Smirnov in 1939. It is a test of goodness-of-fit for continuous data, based on the maximum vertical deviation between the empirical distribution function,  $F_N(x)$ , and the hypothesized cumulative distribution function,  $F_0(x)$ . Small differences support the null hypothesis while large differences are evidence against the null hypothesis.

The null hypothesis is  $H_0: F_N(x) = F_0(x)$  for all  $x$ , and the alternative hypothesis is  $H_1: F_N(x) \neq F_0(x)$  for at least some  $x$  where  $F_0(x)$  is a completely specified continuous distribution. The empirical distribution function,  $F_N(x)$ , is a step function defined as:

$$F_N(x) = \frac{\text{number of sample values} \leq x}{N} \quad (1)$$

where  $N$  = sample size.

*Test statistic.* The test statistic,  $D_N$ , is the maximum vertical distance between the empirical distribution function and the cumulative distribution function.

$$D_N = \max[\max|F_N(x_i) - F_0(x_i)|, \max|F_N(x_{i-1}) - F_0(x_i)|] \quad (2)$$

Both vertical distances  $F_N(x_i) - F_0(x_i)$  and  $F_N(x_{i-1}) - F_0(x_i)$  have to be calculated in order to find the maximum deviation. The overall maximum of the two calculated deviations is defined as  $D_n$ .

For a one-tailed test against the alternatives  $H_1: F_N(x) > F_0(x)$  or  $H_1: F_N(x) < F_0(x)$  for at least some values of  $x$ , the test statistics are respectively:

$$D_N^+ = \max[F_N(x) - F_0(x)] \quad (3)$$

or

$$D_n^- = \max[F_0(x) - F_N(x)] \quad (4)$$

The rejection rule is to reject  $H_0$  when  $D_N \geq D_{N,\alpha}$  where  $D_{N,\alpha}$  is the critical value for sample size  $N$  and level of significance  $\alpha$ .

*Large sample sizes.* The null distribution of  $4ND_N^+$  (or  $4ND_N^-$ ) is approximately  $\chi^2$  with 2 degrees of freedom. Thus, the large sample approximation is

$$D_n^+ \approx \frac{1}{2} \sqrt{\frac{\chi_{\alpha,2}^2}{N}} \quad (5)$$

where  $\chi_{\alpha,2}^2$  is the value for chi-square with 2 degrees of freedom.

*Example.* The K-S goodness-of-fit statistic was calculated for sample 1 (Table 3, Appendix),  $N = 15$ , against the cumulative frequency distribution of the multimodal lumpy data set. The maximum difference at step was 0.07463 and the maximum difference before step was 0.142610. Thus, the value of  $D_n$  is 0.142610. For a two-tail test, with  $\alpha = .05$ , the large sample approximation is

$$1.3581/\sqrt{15} = 1.3581/\sqrt{15} = 0.35066.$$

Because  $0.142610 < 0.35066$ , the null hypothesis cannot be rejected

The Sign Test

The Sign test is credited to Fisher as early as 1925. One of the first papers on the theory and application of the Sign test is attributed to Dixon and Mood in 1946 (Hollander & Wolfe, 1973). According to Neave and Worthington (1988), the logic of the Sign test is “almost certainly the oldest of all formal statistical tests as there is published evidence of its use long ago by J. Arbuthnott (1710)!” (p. 65).

The Sign test is a test for a population median. It can also be used with matched data as a test for equality of medians, specifically when there is only dichotomous data. (Otherwise, the Wilcoxon Signed Rank is more powerful.) The test is based on the number of values above or below the hypothesized median. Gibbons (1971) referred to the Sign test as the nonparametric counterpart of the one-sample  $t$  test. The Sign test tests the null hypothesis  $H_0: M = M_0$ , where  $M$  is the sample median and  $M_0$  is the hypothesized population median, against the alternative hypothesis  $H_1: M \neq M_0$ . One-tailed test alternative hypotheses are of the form  $H_1: M < M_0$  and  $H_1: M > M_0$ .

*Procedure.* Each  $x_i$  is compared with  $M_0$ . If  $x_i > M_0$  then a plus symbol ‘+’ is recorded. If  $x_i < M_0$  then a minus symbol ‘-’ is recorded. In this way all data are reduced to ‘+’ and ‘-’ symbols.

*Test statistic.* The test statistic is the number of ‘+’ symbols or the number of ‘-’ symbols. If the expectation under the alternative hypothesis is that there will be a preponderance of ‘+’ symbols, the test statistic is the number of ‘-’ symbols. Similarly, if the expectation is a preponderance of ‘-’ symbols, the test statistic is the number of ‘+’ symbols. If the test is two-tailed, use the smaller of the two. Thus, depending on the context,

$$S = \text{number of ‘+’ or ‘-’ symbols} \quad (6)$$

*Large sample sizes.* The large sample approximation is given by

$$S^* = \frac{S - \frac{N}{2}}{\sqrt{\frac{N}{4}}} \quad (7)$$

where  $S$  is the test statistic and  $N$  is the sample size.  $S^*$  is compared to the standard normal  $z$  scores for the appropriate  $\alpha$  level.

*Example.* The Sign test was calculated using sample 1 (Table 3, Appendix),  $N = 15$ . The population median is 18.0. The number of minus symbols is 7 and the number of plus symbols is 8. Therefore  $S = 7$ . The large sample approximation,  $S^*$ , using formula (7) is  $-.258199$ . The null hypothesis cannot be rejected because  $-.258199 > -1.95996$ .

Wilcoxon’s Signed Rank Test

The Signed Rank test was introduced by Wilcoxon in 1945. This statistic uses the ranks of the absolute differences between  $x_i$  and  $M_0$  along with the sign of the difference. It uses the relative magnitudes of the data. This statistic can also be used to test for symmetry and to test for equality of location for paired replicates. The null hypothesis is  $H_0: M = M_0$ , which is tested against the alternative  $H_1: M \neq M_0$ . The one-sided alternatives are  $H_1: M < M_0$  and  $H_1: M > M_0$ .

*Procedure.* Compute the differences,  $D_i$ , by the formula

$$D_i = x_i - M_0 \quad (8)$$

Rank the absolute value of the differences in ascending order, keeping track of the individual signs.

*Test statistic.* The test statistic is the sum of either the positive ranks or the negative ranks. If the alternative hypothesis suggests that the sum of the positive ranks should be larger, then

$$T^- = \text{the sum of negative ranks} \quad (9)$$

If the alternative hypothesis suggests that the sum of the negative ranks should be larger, then

$$T^+ = \text{the sum of positive ranks} \quad (10)$$

For a two-tailed test,  $T$  is the smaller of the two rank sums. The total sum of the ranks is  $\frac{N(N+1)}{2}$ , which gives the following relationship:

$$T^+ = \frac{N(N+1)}{2} - T^- \quad (11)$$

*Large sample sizes.* The large sample approximation is given by

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \quad (12)$$

where  $T$  is the test statistic. The resulting  $z$  is compared to the standard normal  $z$  for the appropriate alpha level.

*Example.* The Signed Rank test was computed using the data from sample 1 (Table 3, Appendix),  $N = 15$ . The median of the population is 18.0. Tied differences were assigned midranks. The sum of the negative ranks was 38.5 and the sum of the positive ranks was 81.5. Therefore the Signed Rank statistic is 38.5. The large sample approximation is  $\frac{-21.5}{\sqrt{310}} = \frac{-21.5}{17.6068} = -1.22112$ .

Because  $-1.22112 > -1.95996$ , the null hypothesis is not rejected.

**Estimator of the Median (Continuous Distribution)**

The sample median is a point estimate of the population median. This procedure provides a 1- $\alpha$  confidence interval for the population median. It was designed for continuous data.

*Procedure.* Let  $N$  be the size of the sample. Order the  $N$  observations in ascending order,  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$ . Let  $x_{(0)} = -\infty$  and  $x_{(N+1)} = \infty$ . These  $N+2$  values form  $N+1$  intervals  $(x_{(0)}, x_{(1)}), (x_{(1)}, x_{(2)}), \dots, (x_{(N-1)}, x_{(N)}), (x_{(N)}, x_{(N+1)})$ .

The  $i^{\text{th}}$  interval is defined as  $(x_{(i-1)}, x_{(i)})$  with  $i = 1, 2, \dots, N, N+1$ . The probability that the median is in any one interval can be computed from the binomial distribution. The confidence interval for the median requires that  $r$  be found such that the sum of the probabilities of the intervals in both the lower and upper ends give the best conservative approximation of  $\alpha/2$ , according to the following:

$$\frac{\alpha}{2} \approx \sum_{j=0}^r \binom{N}{j} \frac{1}{2^N} = \sum_{j=N-r}^N \binom{N}{j} \frac{1}{2^N} \quad (13)$$

Thus,  $(x_{(r)}, x_{(r+1)})$  is the last interval in the lower end, making  $x_{(r+1)}$  the lower limit of the confidence

interval. By a similar process,  $x_{(N-r)}$  is the upper limit of the confidence interval.

*Large sample sizes.* Deshpande, Gore, and Shanubhogue (1995) stated “one may use the critical points of the standard normal distribution, to choose the value of  $r + 1$  and  $n - r$ , in the following way”:  $r + 1$  is the integer closest to

$$\frac{N}{2} - z_{\alpha/2} \left( \frac{N}{4} \right)^{\frac{1}{2}} \quad (14)$$

where  $z_{\alpha/2}$  is the upper  $\alpha/2$  critical value of the standard normal distribution.

*Example.* The data from sample 1 (Table 3, Appendix),  $N = 15$ , were used to compute the Estimator of the Median. The population median is 18.0. For the given  $N$  and  $\alpha = .05$ , the value of  $r$  is 3. The value of  $r + 1$  is 4, and  $n - r$  is 12. The 4<sup>th</sup> value is 13 and the 12<sup>th</sup> value is 33. Therefore the interval is (13, 33). The large sample approximation yields  $7.5 - 1.95996(1.9365) = 7.5 - 3.70 = 3.80$ . The closest integer is  $r + 1 = 4$ , so  $r = 3$  and  $N - r = 12$ , resulting in the same interval, (13, 33). The interval contains the population median, 18.0.

**Two Sample Tests**

The two-sample layout consists of independent random samples drawn from two populations. This study examined two sample tests for general differences, two sample location tests, and two sample scale tests.

When differences between two samples are not expected to be predominantly differences in location or differences in scale, a test for general differences is appropriate. Generally differences in variability are related to differences in location. Two tests for differences were considered, the Kolmogorov-Smirnov test for general differences and Rosenbaum’s test.

Two sample location problems involve tests for a difference in location between two samples when the populations are assumed to be similar in shape. The idea is that  $F_1(x) = F_2(x+\theta)$  or  $F_1(x) = F_2(x-\theta)$  where  $\theta$  is the distance between the population medians. Tukey’s quick test, the Wilcoxon (Mann-Whitney) statistic, and the

Hodges-Lehmann estimator of the difference in location for two populations were considered.

In two sample scale tests, the population distributions are usually assumed to have the same location with different spreads. However, Neave and Worthington (1988) cautioned that tests for difference in scale could be severely impaired if there is a difference in location as well. The following nonparametric tests for scale were studied: the Siegel-Tukey test, the Mood test, the Savage test for positive random variables, and the Ansari-Bradley test.

**Kolmogorov-Smirnov Test for General Differences**

The Kolmogorov-Smirnov test compares the cumulative distribution frequencies of the two samples to test for general differences between the populations. The sample cdf “is an approximation of the true cdf of the corresponding population – though, admittedly, a rather crude one if the sample size is small” (Neave & Worthington, 1988, p. 149). This property was used in the goodness-of-fit test above. Large differences in the sample cdfs can indicate a difference in the population cdfs, which could be due to differences in location, spread, or more general differences in the distributions. The null hypothesis is  $H_0 : F_1(x) = F_2(x)$  for all  $x$ . The alternative hypothesis is  $H_1 : F_1(x) \neq F_2(x)$  for some  $x$ .

*Procedure.* The combined observations are ordered from smallest to largest, keeping track of the sample membership. Above each score, write the cdf of sample 1, and below each score write the cdf of sample 2. Because the samples are of equal sizes, it is only necessary to use the numerator of the cdf. For example, the  $cdf(x_i) = \frac{i}{n}$ . Then, write  $i$  above  $x_i$  for sample 1. Find the largest difference between the cdf for sample 1 and the cdf for sample 2.

*Test statistic.* The test statistic is  $D^*$ .  $D^* = n_1 n_2 D$ , and  $D^* = n^2 D$  for equal sample sizes. The above procedure yields  $nD$ . Thus

$$D^* = n(nD) . \tag{15}$$

The greatest difference found by the procedure is multiplied by the sample size.

*Large sample sizes.* The distribution is approximately  $\chi^2$  with 2 degrees of freedom as sample size increases, as it is for the goodness-of-fit test. The large sample approximation for  $D$  is

$$D = \frac{1}{2} \sqrt{\frac{\chi_{\alpha,2}^2 (n_1 + n_2)}{n_1 n_2}} \tag{16}$$

where  $\chi_{\alpha,2}^2$  is the value for chi-square with 2 degrees of freedom for the appropriate alpha level, and  $n_1$  and  $n_2$  are the two sample sizes. The resulting  $D$  is used in formula (15).

*Example.* This example used the data from sample 1 and sample 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The greatest difference ( $nD$ ) between the cdfs of the two samples is  $nD = 3$ . Therefore  $D^* = 15(3) = 45$ . The large sample approximation is  $15^2(1.3581)\sqrt{\frac{30}{225}} = 225(1.3581)(.365148) = 111.579301$ . Because  $45 < 111.579301$ , the null hypothesis cannot be rejected.

**Rosenbaum’s Test**

Rosenbaum’s test, which was developed in 1965, is useful in situations where an increase in the measure of location implies an increase in variation. It is a quick and easy test based on the number of observations in one sample greater than the largest observation in the other sample. The null hypothesis is that both populations have the same location and spread against the alternative, that both populations differ in location and spread.

*Procedure.* The largest observation in each sample is identified. If the largest overall observation is from sample 1, then count the number of observations from sample 1 greater than the largest observation from sample 2. If the largest overall observation is from sample 2, then count the number of observations from sample 2 greater than the largest observation from sample 1.

*Test statistic.* The test statistic is the number of extreme observations.  $R$  is the number of observations from sample 1 greater than the largest observation in sample 2, or the number of observations from sample 2 greater than the largest observation in sample 1.



*Large sample sizes.* As sample sizes increase,  $\frac{n_1}{N} \rightarrow p$  and the probability that the number of extreme values equals  $h$  approaches  $p^h$ .

*Example.* Rosenbaum's statistic was calculated using samples 1 and 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The maximum value from sample 1 is 39, and from sample 2 it is 33. There are three values from sample 1 greater than 33: 34, 36, and 39. Hence,  $R = 3$ . The large sample approximation is  $(.5)^3 = 0.125$ . Because  $0.125 > .05$ , the null hypothesis cannot be rejected.

### Tukey's Quick Test

Tukey published a quick and easy test for the two-sample location layout in 1959. It is easy to calculate and in most cases does not require the use of tables. The most common one-tailed critical values are 6 ( $\alpha = .05$ ) and 9 ( $\alpha = .01$ ). These critical values can be used for most sample sizes. The statistic is the sum of extreme runs in the ordered combined samples. When a difference in location exists, more observations from sample 1 will be expected at one end and more observations from sample 2 will be expected at the other end.

*Procedure.* The combined samples can be ordered, but it is only necessary to order the largest and smallest observations. If both the maximum and minimum values come from the same sample the test is finished, the value of  $T_y = 0$ , and the null hypothesis is not rejected.

For the one-tailed test, the run on the lower end should come from the sample expected to have the lower median, and the run on the upper end should come from the sample expected to have the larger median. For a two-tailed test, it is possible to proceed with the test as long as the maximum and minimum observations come from different samples.

*Test statistic.*  $T_y$  is defined as follows for the alternative hypothesis,  $H_1: M_1 > M_2$ .  $T_y$  is the number of observations from sample 2 less than the smallest observation of sample 1, plus the number of observations from sample 1 greater than the largest observation from sample 2. For the alternative  $H_1: M_2 > M_1$  the samples are reversed. For the two-tailed hypothesis  $H_1: M_1 \neq M_2$ , both possibilities are considered.

*Critical values.* As stated above, generally, the critical value for  $\alpha = .05$  is 6, and is 9 for  $\alpha =$

.01. There are tables available. As long as the ratio of  $n_x$  to  $n_y$  is within 1 to 1.5, these critical values work well. There are corrections available when the ratio exceeds 1.5. For a two-tailed test the critical values are 7 ( $\alpha = .05$ ) and 10 ( $\alpha = .01$ ).

*Large sample sizes.* The null distribution is based on the order of the elements of both samples at the extreme ends. It does not depend on the order of the elements in the middle. Neave and Worthington (1988, p. 125) gave the following formula:

$$\text{Prob}(T_y \geq h) = \frac{pq(q^h - p^h)}{q - p} \quad (17)$$

for  $h \geq 2$ . When the sample sizes are equal,  $p = q = .5$ . Then the probability of  $T_y \geq h$  is  $h2^{-(h+1)}$ . For a two-tailed test the probability is doubled.

*Example.* The Tukey test was calculated using the data in sample 1 and sample 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The maximum value (39) is from sample 1 and the minimum (2) is from sample 5, so the test may proceed. The value of  $T_y = 1 + 3 = 4$ . For a two-tailed test with  $\alpha = .05$ , the large sample approximation is  $2(4)(2^{-5}) = 0.25$ . Because  $0.25 > .05$ , the null hypothesis cannot be rejected.

### Wilcoxon (Mann-Whitney) Test

In 1945, Wilcoxon introduced the Rank Sum test, and in 1947 Mann and Whitney presented a different version of the same test. The Wilcoxon statistic is easily converted to the Mann-Whitney  $U$  statistic. The hypotheses of the test are  $H_0: F_1(x) = F_2(x)$  for all  $x$  against the two-tailed alternative,  $H_0: F_1(x) \neq F_2(x)$ . The one-tailed alternative is  $H_1: F_1(x) = F_2(x + \theta)$ .

*Procedure.* For the Wilcoxon test, the combined samples are ordered, keeping track of sample membership. The ranks of the sample that is expected, under the alternative hypothesis, to have the smallest sum, are added. The Mann-Whitney test is conducted as follows. Put all the observations in order, noting sample membership. Count how many of the observations of one sample exceed each observation in the first sample. The sum of these counts is the test statistic,  $U$ .

*Test statistic.* For the Wilcoxon test,

$$S_n = \sum_{j=1}^n R_j \tag{18}$$

where  $R_j$  are the ranks of sample  $n$  and  $S_n$  is the sum of the ranks of the sample expected to have the smaller sum.

For the Mann-Whitney test, calculate the  $U$  statistic for the sample expected to have the smaller sum under the alternative hypothesis.

$$U_{n_2} = \text{the sum of the observations in } n_1 \text{ exceeding each observation in } n_2. \tag{19}$$

$$U_{n_1} = \text{the sum of the observations in } n_2 \text{ exceeding each observation in } n_1. \tag{20}$$

There is a linear relation between  $S_n$  and  $U_n$ . It is expressed as

$$U_{n_1} = S_{n_1} - \frac{1}{2}n_1(n_1 + 1) \tag{21}$$

and similarly,

$$U_{n_2} = S_{n_2} - \frac{1}{2}n_2(n_2 + 1) \tag{22}$$

where

$$U_{n_1} = n_1n_2 - U_{n_2} \tag{23}$$

In a two-tailed test, use the smallest  $U$  statistic to test for significance.

*Large sample sizes.* The large-sample approximation using the Wilcoxon statistic,  $S_{n_1}$  is:

$$z = \frac{S_{n_1} - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1n_2(n_1 + n_2 + 1)}{12}}} \tag{24}$$

The large-sample approximation with the  $U$  statistic is

$$z = \frac{U + \frac{1}{2} - \frac{1}{2}n_1n_2}{\sqrt{\frac{n_1n_2(n_1 + n_2 + 1)}{12}}} \tag{25}$$

In either case, reject  $H_0$  if  $z < -z_\alpha$  (or  $z < -z_{\alpha/2}$  for a two-tailed test).

*Example.* The Wilcoxon Rank Sum (Mann-Whitney) statistic was calculated with data from sample 1 and sample 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The combined samples were ranked, using midranks in place of the ranks of tied observations. The rank sum for sample 1 was 258.5 and for sample 5, 206.5. Hence  $S = 206.5$ . Calculating the  $U$  statistic,  $U = 206.5 - 0.5(15)(16) = 86.5$ . The large sample approximation for  $U$  is  $\frac{86.5 + .5 - .5(15^2)}{\sqrt{\frac{15^2(31)}{12}}} = \frac{-25.5}{24.1091} = -1.05769$ . Because  $-1.05769 > -1.95996$ , the null hypothesis cannot be rejected.

Hodges-Lehmann Estimator of the Difference in Location

It is often useful to estimate the difference in location between two populations. Suppose two populations are assumed to have similar shapes, but differ in locations. The objective is to develop a confidence interval that will have the probability of  $1-\alpha$  that the difference lies within the interval.

*Procedure.* All the pairwise differences are computed,  $x_i - y_j$ . For sample sizes of  $n_1$  and  $n_2$ , there are  $n_1n_2$  differences. The differences are put in ascending order. The task is to find two integers  $l$  and  $u$  such that the probability that the difference lies between  $l$  and  $u$  is equal to  $1-\alpha$ . These limits are chosen symmetrically. The appropriate lower tail critical value is found for the Mann-Whitney  $U$  statistic. This value is the upper limit of the lower end of the differences. Therefore,  $l$  is the next consecutive integer. The upper limit of the confidence interval is the  $u^{\text{th}}$  difference from the upper end, found by  $u = n_1n_2 - l + 1$ . The interval  $(l, u)$  is the confidence interval for the difference in location for the two populations.

*Large sample sizes.* Approximate  $l$  and  $u$  by

$$l = \left[ \frac{n_1 n_2}{2} - z_{\alpha/2} \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} - \frac{1}{2} \right] \quad (26)$$

and

$$u = \left[ \frac{n_1 n_2}{2} + z_{\alpha/2} \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} - \frac{1}{2} \right] \quad (27)$$

“where the square brackets denote integer nearest to the quantity within, and  $z_{\alpha/2}$  is the suitable upper critical point of the standard normal distribution” (Deshpande, et al., 1995, p. 45, formulas rewritten for consistency of notation with this article).

*Example.* The Hodges-Lehmann estimate of the difference in location was computed using samples 1 and 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . All possible differences were computed and ranked. Using the large sample approximation formula (26),  $l = 112.5 - 1.95596(24.109) - 0.5 = 64.844$ . Thus,  $l = 65$  and the lower bound is the 65<sup>th</sup> difference, which is -4. The upper bound is the 65<sup>th</sup> difference from the upper end, or the 225 - 65 + 1 = 161<sup>st</sup> value, 14. The confidence interval is (-4, 14).

### Siegel-Tukey Test

The Siegel-Tukey test was developed in 1960. It is similar in procedure to the Wilcoxon Rank Sum test for difference in location. It is based on the logic that if two samples come from populations with the same median, the one with the greater variability will have more extreme scores. An advantage of the Siegel-Tukey statistic is that it uses the Wilcoxon table of critical values or can be transformed into a  $U$  statistic for use with the Mann-Whitney  $U$  table of critical values.

The hypotheses for a two-tailed test are  $H_0$ : There is no difference in spread between the two populations, which is tested against the alternative  $H_1$ : There is some difference in spread between the two populations.

*Procedure.* The two combined samples are ordered, keeping track of sample membership. The ranking proceeds as follows: the lowest observation is ranked 1, the highest is ranked 2, and the next highest 3. Then the second lowest is

ranked 4 and the subsequent observation ranked 5. The ranking continues to alternate from lowest to highest, ranking two scores at each end. If there is an odd number of scores, the middle score is discarded and the sample size reduced accordingly. Below is an illustration of the ranking procedure:

1 4 5 8 9 ... N ... 7 6 3 2  
 where  $N = n_1 + n_2$ .

*Test statistic.* The sum of ranks is calculated for one sample. The rank sum can be used with a table of critical values or it can be transformed into a  $U$  statistic by one of the following formulas:

$$U^* = R_{n_1} - \frac{1}{2} n_1 (n_1 + 1) \quad (28)$$

or

$$U^* = R_{n_2} - \frac{1}{2} n_2 (n_2 + 1). \quad (29)$$

*Large sample sizes.* The large-sample approximations are the same for the Siegel-Tukey test as for the Wilcoxon Rank Sum or the Mann-Whitney  $U$  statistic, formulas (24) and (25).

*Example.* The Siegel-Tukey statistic was calculated using sample 1 and sample 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The samples were combined and ranked according to the method described. Then, tied ranks were averaged. The sum of ranks was 220.5 for sample 1, and 244.5 for sample 5. The  $U$  statistic is  $220.5 - .5(15)(16) = 100.5$ . The large sample approximation is

$$z = \frac{100.5 + .5 - .5(15^2)}{\sqrt{\frac{15^2(31)}{12}}} = \frac{-11.5}{24.109127} = -0.476998.$$

Because  $-0.476998 > -1.95996$ , the null hypothesis cannot be rejected.

### The Mood Test

In 1954, the Mood test was developed based on the sum of squared deviations of one sample's ranks from the average combined ranks. The null hypothesis is that there is no difference in spread against the alternative hypothesis that there is some difference.

*Procedure.* Let sample 1 be  $x_1, x_2, \dots, x_{n_1}$  and let sample 2 be  $y_1, y_2, \dots, y_{n_2}$ . Arrange the combined samples in ascending order and rank the observations from 1 to  $n_1 + n_2$ . Let  $R_i$  be the rank of  $x_i$ . Let  $N = n_1 + n_2$ . If  $N$  is odd, the middle rank is ignored to preserve symmetry.

*Test statistic.* The test statistic is

$$M = \sum_{i=1}^{n_1} \left( R_i - \frac{n_1 + n_2 + 1}{2} \right)^2. \quad (30)$$

Large sample sizes. The large sample approximation is

$$z = \frac{M - \frac{n_1(N^2 - 1)}{12}}{\sqrt{\frac{n_1 n_2 (N + 1)(N^2 - 4)}{180}}} \quad (31)$$

where  $N = n_1 + n_2$  and  $M$  is the test statistic.

*Example.* The Mood statistic was calculated using sample 1 and sample 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The combined samples are ranked, with midranks assigned to the ranks of tied observations. The mean of the ranks is 15.5, and the sum of squared deviations of the ranks from the mean for sample 1 was calculated, yielding  $M=1257$ . The large sample approximation is  $\frac{1257 - 1123.75}{\sqrt{34720}} = \frac{133.25}{186.333} = 0.71512$ . Because  $0.71512 < 1.95596$ , the null hypothesis cannot be rejected.

The Savage Test for Positive Random Variables

Unlike the Siegel-Tukey test and the Mood test, the Savage test does not assume that location remains the same. It is assumed that differences in scale cause a difference in location. The samples are assumed to be drawn from continuous distributions.

The null hypothesis is that there is no difference in spread, which is tested against the two-tailed alternative that there is a difference in variability.

*Procedure.* Let sample 1 be  $x_1, x_2, \dots, x_{n_1}$  and let sample 2 be  $y_1, y_2, \dots, y_{n_2}$ . The combined samples are ordered, keeping track of sample

membership. Let  $R_i$  be the rank for  $x_i$ . The test statistic is computed for either sample.

*Test statistic.* The test statistic is

$$S = \sum_{i=1}^{n_1} a(R_i) \quad (32)$$

where

$$a(i) = \sum_{j=N+1-i}^N \frac{1}{j} \quad (33)$$

such that

$$a(1) = \frac{1}{N}, a(2) = \frac{1}{N-1} + \frac{1}{N}, \dots, a(N) = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{N-1} + \frac{1}{N}$$

*Large sample sizes.* For large sample sizes the following normal approximation may be used.

$$S^* = \frac{S - n_2}{\sqrt{\frac{n_1 n_2}{N-1} \left( 1 - \frac{1}{N} \sum_{j=1}^N \frac{1}{j} \right)}} \quad (34)$$

$S^*$  is compared to the critical  $z$  value from the standard normal distribution.

*Example.* The Savage statistic was calculated using samples 1 and 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . Using sample 1,  $S = 18.3114$ . The large sample approximation is  $\frac{18.3114 - 15}{\sqrt{7.7586(0.86683)}} = \frac{3.114}{2.59334} = 1.27689$ . Because  $1.27689 < 1.95596$ , the null hypothesis cannot be rejected.

Ansari-Bradley Test

This is a rank test for spread when the population medians are the same. The null hypothesis is that the two populations have the same spread, which is tested against the alternative that the variability of the two populations differs.

*Procedure.* Order the combined samples, keeping track of sample membership. Rank the smallest and largest observation 1. Rank the second lowest and second highest 2. If the combined sample size,  $N$ , is odd, the middle score will be ranked  $\frac{N+1}{2}$  and if  $N$  is even the middle

two ranks will be  $\frac{N}{2}$ . The pattern will be either 1, 2, 3, . . . ,  $\frac{N+1}{2}$ , . . . , 3, 2, 1 (N odd), or 1, 2, 3, . . . ,  $\frac{N}{2}$ ,  $\frac{N}{2}$ , . . . , 3, 2, 1 (N even).

*Test statistic.* The test statistic, W, is the sum of the ranks of sample 1.

$$W = \sum_{i=1}^{n_1} R_i \tag{35}$$

where  $R_i$  is the rank of the  $i^{\text{th}}$  observation of a sample.

*Large sample sizes.* There are two formulas. If N is even, use

$$W^* = \frac{W - \frac{n_1(n_1 + n_2 + 2)}{4}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 2)(n_1 + n_2 - 2)}{48(n_1 + n_2 - 1)}}} \tag{36}$$

and if N is odd, use

$$W^* = \frac{W - \frac{n_1(n_1 + n_2 + 1)^2}{4(n_1 + n_2)}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)[3 + (n_1 + n_2)^2]}{48(n_1 + n_2)^2}}} \tag{37}$$

Reject the null hypothesis if  $W^* \geq z_{\alpha/2}$ .

*Example.* The Ansari-Bradley statistic was calculated using samples 1 and 5 (Table 3, Appendix),  $n_1 = n_2 = 15$ . The combined samples were ranked using the method described, and the ranks of tied observations were assigned average ranks. The two-tailed statistic, W, is 126.5, the rank sum of sample 5. The large sample approximation is  $\frac{126.5 - 120}{\sqrt{144.8276}} = \frac{6.5}{12.034} = 0.54$ .

Because  $0.54 < 1.95596$ , the null hypothesis cannot be rejected.

### Comparisons Of Several Populations

This section considered tests against an omnibus alternative and tests involving an ordered hypothesis. The omnibus tests were the Kruskal-Wallis test and Friedman’s test. The tests for

ordered alternatives are the Terpstra-Jonckheere test, Page’s test, and the Match test.

The Kruskal-Wallis statistic is a test for independent samples. It is analogous to the one-way analysis of variance. Friedman’s test is an omnibus test for k related samples, and is analogous to a two-way analysis of variance.

Comparisons of several populations with ordered alternative hypotheses are extensions of a one-sided test. When an omnibus alternative states only that there is some difference between the populations, an ordered alternative specifies the order of differences. Three tests for an ordered alternative were included: the Terpstra-Jonckheere Test, Page’s Test, and the Match Test.

### Kruskal-Wallis Test

The Kruskal-Wallis test was derived from the F test in 1952. It is an extension of the Wilcoxon (Mann-Whitney) test. The null hypothesis is that the k populations have the same median. The alternative hypothesis is that at least one sample is from a distribution with a different median.

*Procedure.* Rank all the observations in the combined samples, keeping track of the sample membership. Compute the rank sums of each sample. Let  $R_i$  equal the sum of the ranks of the  $i^{\text{th}}$  sample of sample size  $n_i$ . The logic of the test is that the ranks should be randomly distributed among the k samples.

*Test statistic.* The formula is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \tag{38}$$

where N is the total sample size,  $n_i$  is the size of the  $i^{\text{th}}$  group, k is the number of groups, and  $R_i$  is the rank-sum of the  $i^{\text{th}}$  group. Reject  $H_0$  when  $H \geq$  critical value.

*Large sample sizes.* For large sample sizes, the null distribution is approximated by the  $\chi^2$  distribution with  $k - 1$  degrees of freedom. Thus, the rejection rule is to reject  $H_0$  if  $H \geq \chi_{\alpha, k-1}^2$  where  $\chi_{\alpha, k-1}^2$  is the value of  $\chi^2$  at nominal  $\alpha$  with  $k - 1$  degrees of freedom.

*Example.* The Kruskal-Wallis statistic was calculated using samples 1–5 (Table 3, Appendix),  $n_1 = n_2 = n_3 = n_4 = n_5 = 15$ . The combined samples

were ranked, and tied ranks were assigned midranks. The rank sums were:  $R_1 = 638$ ,  $R_2 = 595$ ,  $R_3 = 441.5$ ,  $R_4 = 656.5$ , and  $R_5 = 519$ . The sum of  $R_i^2 = 1,656,344.5$ ,  $i = 1, 2, 3, 4, 5$ .

$$H = \frac{12}{75(76)} \left( \frac{1,656,344.5}{15} \right) - 3(76) = 0.00211(110,422.97 - 228) = 4.47$$

Thus,  $H = 4.47$ . The large sample approximation with  $5 - 1 = 4$  degrees of freedom at  $\alpha = .05$  is  $\chi^2 = 9.488$ . Because  $4.47 < 9.488$ , the null hypothesis cannot be rejected.

**Friedman's Test**

The Friedman test was developed as a test for  $k$  related samples in 1937. The null hypothesis is that the samples come from the same population. The alternative hypothesis is that at least one of the samples comes from a different population. Under the truth of the null hypothesis, this test only requires exchangeability (or, if variances differ, compound symmetry) and the ability to rank the data. The data are arranged in  $k$  columns and  $n$  rows, where each row contains  $k$  related observations.

*Procedure.* Rank the observations for each row from 1 to  $k$ . For each of the  $k$  columns, the ranks are added and averaged, and the mean is designated  $\bar{R}_j$ . The overall mean of the ranks is  $\bar{R} = \frac{1}{2}(k+1)$ . The sum of the squares of the deviations of mean of the ranks of the columns from the overall mean rank is computed. The test statistic is a multiple of this sum.

*Test statistic.* The test statistic for Friedman's test is  $M$ , which is a multiple of  $S$ , as follows:

$$S = \sum_{j=1}^k (\bar{R}_j - \bar{R})^2 \tag{39}$$

$$M = \frac{12n}{k(k+1)} S \tag{40}$$

where  $n$  is the number of rows, and  $k$  is the number of columns. An alternate formula that does not use  $S$  is as follows.

$$M = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1) \tag{41}$$

where  $n$  is the number of rows,  $k$  is the number of columns, and  $R_j$  is the rank sum for the  $j^{\text{th}}$  column,  $j = 1, 2, 3, \dots, k$ .

*Large sample sizes.* For large sample sizes, the critical values can be approximated by  $\chi^2$  with  $k - 1$  degrees of freedom.

*Example.* Friedman's statistic was calculated with samples 1 - 5 (Table 3, Appendix),  $n_1 = n_2 = n_3 = n_4 = n_5 = 15$ . The rows were ranked, with the ranks of tied observations replaced with midranks. The column sums are:  $R_1 = 48.5$ ,  $R_2 = 47$ ,  $R_3 = 33$ ,  $R_4 = 52.5$ , and  $R_5 = 44$ . The sum of the squared rank sums is 10,342.5.  $M = \frac{12}{15 \cdot 5 \cdot 6} (10,342.5) - 3 \cdot 15 \cdot 6 = 0.0267(10,342.5) - 270 = 5.8$ . The large sample approximation is  $\chi^2$  with  $5 - 1 = 4$  degrees of freedom and  $\alpha = .05$ , which is 9.488. Because  $5.8 < 9.488$ , the null hypothesis cannot be rejected.

**Terpstra-Jonckheere Test**

This is a test for more than two independent samples. It was first developed by Terpstra in 1952 and later independently developed by Jonckheere in 1954. The null hypothesis is that the medians of the samples are equal, which is tested against the alternative that the medians are either decreasing or increasing. This test is based on the Mann-Whitney U statistic, where  $U$  is calculated for each pair of samples and the  $U$  statistics are added.

Suppose the null hypothesis is  $H_0: F_1(x) \geq F_2(x) \geq F_3(x) \geq \dots \geq F_k(x)$  and the alternative hypothesis is  $H_0: F_1(x) < F_2(x) < F_3(x) < \dots < F_k(x)$  for  $i = 1, 2, \dots, k$ . The  $U$  statistic is calculated for each of the  $\frac{k(k-1)}{2}$  pairs, which are ordered so that the smallest  $U$  is calculated.

*Test statistic.* The test statistic is the sum of the  $U$  statistics.

$$W = U_{k,1} + U_{k,2} + \dots + U_{3,1} + U_{3,2} + U_{2,1} \tag{42}$$

where  $U_{ij}$  is the number of pairs when the observation from sample  $j$  is less than the observation from sample  $i$ .

Large sample sizes. The null distribution of  $W$  approaches normality as the sample size increases. The mean of the distribution is

$$\mu = \frac{(N^2 - \sum n_i^2)}{4} \tag{43}$$

and the standard deviation is

$$\sigma = \sqrt{\frac{N^2(2N+3) - \sum n_i^2(2n_i+3)}{72}} \tag{44}$$

The critical value for large samples is given by

$$W \leq \mu - z\sigma - \frac{1}{2} \tag{45}$$

where  $z$  is the standard normal value, and  $\frac{1}{2}$  is a continuity correction.

*Example.* The Terpstra-Jonckheere statistic was calculated with samples 1 – 5 (Table 3, Appendix),  $n_1 = n_2 = n_3 = n_4 = n_5 = 15$ . This was done as a one-tailed test with  $\alpha = .05$ . The  $U$  statistics for each sample were calculated.  $U_{5,1} = 135$ ,  $U_{5,2} = 124$ ,  $U_{5,3} = 91$ ,  $U_{5,4} = 136$ ,  $U_{4,1} = 103$ ,  $U_{4,2} = 97$ ,  $U_{4,3} = 71$ ,  $U_{3,1} = 145$ ,  $U_{3,2} = 142$ , and  $U_{2,1} = 121$ , for a total  $W = 1,165$ . The large sample approximation was calculated with  $\mu = 1125$  and  $\sigma = 106.94625$ . The approximation is  $1125 - 1.6449(106.9463) - .5 = 948.584$ . Because  $1165 > 948.584$  the null hypothesis cannot be rejected.

Page’s Test

Page’s test for an ordered hypothesis for  $k > 2$  related samples was developed in 1963. It takes the form of a randomized block design with  $k$  columns and  $n$  rows. The null hypothesis is  $H_0 : M_1 = M_2 = \dots = M_k$  and the alternative hypothesis is  $H_1 : M_1 < M_2 < \dots < M_k$  for  $i = 1, 2, \dots k$ . For this test, the alternative must be of this form. The samples need to be reordered if necessary.

*Procedure.* The data are ranked from 1 to  $k$  for each row, creating a table of the ranks. The ranks of each of the  $k$  columns are totaled. If the null hypothesis is true, the ranks should be evenly distributed over the columns, whereas if the

alternative is true, the ranks sums should increase with the column index.

*Test statistic.* Each column rank-sum is multiplied by the column index. The test statistic is

$$L = \sum_{i=1}^k iR_i \tag{46}$$

where  $i$  is the column index,  $i = 1, 2, 3, \dots, k$ , and  $R_i$  is the rank sum for the  $i^{\text{th}}$  column.

Large sample sizes. The mean of  $L$  is

$$\mu = \frac{nk(k+1)^2}{4} \tag{47}$$

and the standard deviation is

$$\sigma = \sqrt{\frac{nk^2(k+1)(k^2-1)}{144}} \tag{48}$$

For a given  $\alpha$ , the approximate critical region is

$$L \geq \mu + z\sigma + \frac{1}{2} \tag{49}$$

*Example.* Page’s statistic was calculated with samples 1 – 5 (Table 3, Appendix),  $n_1 = n_2 = n_3 = n_4 = n_5 = 15$ . This was done as a one-tailed test with  $\alpha = .05$ . The rows are ranked with midranks assigned to tied ranks. The column sums are:  $R_1 = 48.5$ ,  $R_2 = 47$ ,  $R_3 = 33$ ,  $R_4 = 52.5$ , and  $R_5 = 44$ . The statistic,  $L$ , is the sum of  $iR_i^2 = 671.5$ , where  $i = 1, 2, 3, 4, 5$ . The large sample approximation was calculated with  $\mu = 675$  and  $\sigma = 19.3649$ . The approximation is  $675 + 1.64485(19.3649) + .5 = 707.352$ . Because  $671.5 < 707.352$ , the null hypothesis cannot be rejected.

The Match Test for Ordered Alternatives

The Match test is a test for  $k > 2$  related samples with an ordered alternative hypothesis. The Match test was developed by Neave and Worthington (1988). It is very similar in concept to Page’s test, but instead of using rank-sums, it uses the number of matches of the ranks with the expected ranks plus half the near matches. The

null hypothesis is  $H_0: M_1 = M_2 = \dots = M_k$  and the alternative hypothesis is  $H_0: M_1 < M_2 < \dots < M_k$  for  $i = 1, 2, \dots, k$ .

*Procedure.* A table of ranks is compiled with the observations in each row ranked from 1 to  $k$ . Tied observations are assigned average ranks. Each rank,  $r_i$ , is compared with the expected rank,  $i$ , the column index. If the rank equals the column index, it is a match. Count the number of matches. Every non-match such that  $0.5 \leq |r_i - i| \leq 1.5$  is counted as a near match.

*Test statistic.* The test statistic is

$$L_2 = L_1 + \frac{1}{2}(\text{number of near matches}) \quad (50)$$

where  $L_1$  is the number of matches.

Large sample sizes. The null distribution approaches a normal distribution for large sample size. The mean and standard deviation for  $L_2$  are as follows:

$$\mu = n \left( 2 - \frac{1}{k} \right) \quad (51)$$

and

$$\sigma = \sqrt{\frac{n}{k} \left( \frac{3(k-2)}{2} \right) + \frac{1}{k(k-1)}} \quad (52)$$

For a given level of significance  $\alpha$  the critical value approximation is

$$L_2 \geq \mu + z\sigma + \frac{1}{2} \quad (53)$$

where  $z$  is the upper-tail critical value from the standard normal distribution and  $\frac{1}{2}$  is a continuity correction.

*Example.* The Match statistic was calculated with samples 1 – 5 (Table 3, Appendix),  $n_1 = n_2 = n_3 = n_4 = n_5 = 15$ . This was done as a one-tailed test with  $\alpha = .05$ . The rows are ranked, with midranks assigned for tied observations. The number of matches for the five columns are 3, 3, 2, 2, and 1, for  $L_1 = 11$ . The number of near matches were 1, 6, 8, 8, and 4, for  $L_2 = 27$ . The

statistic,  $L = 11 + .5(27) = 24.5$ . For the large sample approximation,  $\mu = 27$  and  $\sigma = 3.68103$ . The approximation is  $27 + 1.6449(3.68103) + .5 = 33.5549$ . Because  $24.5 < 33.5549$ , the null hypothesis cannot be rejected.

Rank Correlation Tests

The rank correlation is a measure of the association of a pair of variables. Spearman’s rank correlation coefficient (rho) and Kendall’s rank correlation coefficient (tau) were studied.

Spearman’s Rank Correlation Coefficient

Spearman’s rank correlation (rho) was published in 1904. Let  $X$  and  $Y$  be the two variables of interests. Each observed pair is denoted  $(x_i, y_i)$ . The paired ranks are denoted  $(r_i, s_i)$ , where  $r_i$  is the rank of  $x_i$  and  $s_i$  is the rank of  $y_i$ . The null hypothesis for a two-tailed test is  $H_0: \rho = 0$ , which is tested against the alternative  $H_1: \rho \neq 0$ . The alternative hypotheses for a one-tailed test are  $H_1: \rho > 0$  or  $H_1: \rho < 0$ .

*Procedure.* Rank both  $X$  and  $Y$  scores while keeping track of the original pairs. Form the rank pairs  $(r_i, s_i)$  which correspond to the original pair,  $(x_i, y_i)$ . Calculate the sum of the squared differences between  $r_i$  and  $s_i$ .

Test statistic. If there are no ties, the formula is

$$\rho = 1 - \frac{6T}{n(n^2 - 1)} \quad (54)$$

where

$$T = \sum (r_i - s_i)^2 \quad (55)$$

Large sample sizes. For large  $n$  the distribution of  $\rho$  is approximately normal. The critical values can be found by  $z = \rho\sqrt{n-1}$ . The rejection rule for a two-tailed test is to reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$  where  $z_{\alpha/2}$  is the critical value for the given level of significance.

*Example.* Spearman’s rho was calculated using sample 1 and sample 5 (Table 3, Appendix),  $n = 15$ . The sum of the squared rank differences for the two samples is  $T = 839$ . Rho is  $1 - \frac{6(839)}{15(224)} = 1 - \frac{5034}{3360} = 1 - 1.498 = -0.498$ . So  $z =$



$-0.498\sqrt{14} = -1.864$ . Because  $-1.864 > -1.956$ , the null hypothesis cannot be rejected.

#### Kendall's Rank Correlation Coefficient

Kendall's rank correlation coefficient ( $\tau$ ) is similar to Spearman's  $\rho$ . The underlying concept is the tendency for concordance, which means that if  $x_i > x_j$  then  $y_i > y_j$ . Concordance implies that the differences  $x_i - x_j$  and  $y_i - y_j$  have the same sign, either "+" or "-". Discordant pairs have opposite signs, that is,  $x_i > x_j$  but  $y_i < y_j$ , or the opposite,  $x_i < x_j$  but  $y_i > y_j$ .

*Procedure.* Arrange the pairs in ascending order of  $X$ . Count the number of  $y_i$  smaller than  $y_1$ . This is the number of disconcordant pairs ( $N_D$ ) for  $x_1$ . Repeat the process for each  $x_i$ , counting the number of  $y_j < y_i$ , where  $j = i + 1, i + 2, i + 3, \dots, n$ .

*Test statistic.* Because the total number of pairs is  $\frac{1}{2}n(n-1)$ ,  $N_c = \frac{1}{2}n(n-1) - N_D$ . The tau statistic ( $\tau$ ) is defined as

$$\tau = \frac{N_c - N_D}{\frac{1}{2}n(n-1)} \quad (56)$$

This formula can be simplified by substituting  $N_c = \frac{1}{2}n(n-1) - N_D$  into the formula so that

$$\tau = 1 - \frac{4N_D}{n(n-1)} \quad (57)$$

*Large sample sizes.* For large sample sizes, the formula is

$$z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \quad (58)$$

where  $z$  is compared to the  $z$  score from the standard normal distribution for the appropriate alpha level.

*Example.* Kendall's tau was calculated using sample 1 and sample 5 (Table 3, Appendix),  $n = 15$ . The number of discordant pairs for each pair,  $(x_1, x_5)$ , were 12, 8, 8, 5, 9, 5, 6, 3, 5, 3, 0, 3,

0, 1, and 0. The total number of discordant pairs,  $N_D$  is 68. Tau is  $1 - \frac{4 \cdot 68}{15 \cdot 14} = 1 - \frac{272}{210} = -0.295$ .

$$\text{Thus } z = \frac{3(-.295)\sqrt{(15)(14)}}{\sqrt{2(35)}} = \frac{-12.835}{8.366} = -1.534.$$

Because  $-1.534 > -1.95596$ , the null hypothesis cannot be rejected.

#### References<sup>1</sup>

\*Anderson, D.R., Sweeney, D.J., & Williams, T. A. (1999). *Statistics for business and economics* (7<sup>th</sup> ed.). Cincinnati: South-Western College Publishing Co.

\*Berenson, M. L., Levine, D. M., & Rindskopf, D. (1988). *Applied statistics: A first course*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *The American Statistician*, 54, 72-77.

Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples rank transformation to that of Wilcoxon's signed rank statistic. *Journal of Educational Statistics*, 10, 368-383.

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's  $t$  statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5, 309-335.

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transformation in factorial ANOVA. *Communications in Statistics*, 16, 1133-1145.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.

Bradstreet, T. E. (1997). A Monte Carlo study of Type I error rates for the two-sample Behrens-Fisher problem with and without rank transformation. *Computational Statistics and Data Analysis*, 25, 167-179.

- Bridge, P. K., & Sawilowsky, S. S. (1999). Increasing physician's awareness of the impact of statistical tests on research outcomes: Investigating the comparative power of the Wilcoxon Rank-Sum test and independent samples t test to violations from normality. *Journal of Clinical Epidemiology*, 52, 229-235.
- Conover, W. J. (1971). *Practical Nonparametric statistics*. New York: John Wiley & Sons, Inc.
- \*Daly, F., Hand, D.J., Jones, M.C., Lunn, A..D., & McConway, K.J. (1995). *Elements of statistics*. Workingham, England: Addison-Wesley.
- \*Daniel, W.W. (1978). *Applied nonparametric statistics*. Boston: Houghton Mifflin Co.
- Deshpande, J.V., Gore, A.P., & Shanubhogue, A.. (1995). *Statistical analysis of nonnormal data*. New York: John Wiley & Sons, Inc.
- \*Ferguson, G. A. (1971). *Statistical analysis in psychology and education* (3<sup>rd</sup> ed.). New York: McGraw-Hill book Company.
- \*Ferguson, G. A. (1981). *Statistical analysis in psychology and education* (5<sup>th</sup> ed.). New York: McGraw-Hill Book Company.
- \*Gravetter, F. J., & Wallnau, L. B. (1985). *Statistics for the behavioral sciences*. St. Paul: West Publishing Co.
- Gibbons, J. D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill Book Company.
- Hájek, J. (1969). *A course in nonparametric statistics*. San Francisco: Holden-Day.
- Harwell, M., & Serlin, R. C. (1997). An empirical study of five multivariate tests for the single-factor repeated measures model. *Communications in Statistics*, 26, 605-618.
- \*Hays, W. L. (1994). *Statistics* (5<sup>th</sup> ed.). Fort Worth: Harcourt Brace College Publishers.
- Headrick, T. C., & Sawilowsky, S. S. (2000). Type I error and power of the RT ANCOVA. American Educational Research Association, SIG/Educational Statisticians. New Orleans, LA
- Headrick, T. C., & Sawilowsky, S. S. (1999). Type I error and power of the rank transform in factorial ANCOVA. Statistics Symposium on Selected Topics in Nonparametric Statistics. Gainesville, FL.
- \*Hildebrand, D. (1986). *Statistical thinking for behavioral scientists*. Boston: Duxbury Press.
- Hollander, M. & Wolfe, D. (1973). *Nonparametrical statistical methods*. New York: John Wiley & Sons.
- \*Jarrett, J. & Kraft, A. (1989). *Statistical analysis for decision making*. Boston: Allyn and Bacon.
- Jonckheere, A. R. (1954). A distribution-free *k*-sample test against ordered alternatives. *Biometrika*, 41, 133-143.
- Kelley, D. L., & Sawilowsky, S. S. (1997). Nonparametric alternatives to the F statistic in analysis of variance. *Journal of Statistical Computation and Simulation*, 58, 343-359.
- \*Knoke, D. and Bohrnstedt, G. W. (1991). *Basic social statistics*. New York: F. E. Peacock Publishers, Inc.
- \*Kraft, C. H. & van Eeden, C. (1968). *A nonparametric introduction to statistics*. New York: Macmillan Co.
- \*Krauth, J. (1988). *Distribution-free statistics: An application-oriented approach*. Amsterdam: Elsevier.
- \*Kurtz, N. R. (1983). *Introduction to social statistics*. New York: McGraw-Hill Book Co.
- Lahey (1998). *Essential Lahey Fortran 90*. Incline Village, NY: Lahey Computer Systems, Inc.
- \*Lehmann, E. L. & D'Abrera, H.J.M. (1975). *Nonparametric statistical methods based on ranks*. New York: McGraw-Hill International Book Company.
- Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *The American Statistician*, 52, 127-132.
- \*Manoukian, E. B. (1986). *Mathematical nonparametric statistics*. New York: Gordon & Breach Science Publications.
- \*McClave, J. T., Dietrich II, F. H. (1988). *Statistics* (4<sup>th</sup> ed.). San Francisco: Dellen Publishing Company.
- \*Mendenhall, W. & Reinmuth, J. E. (1978). *Statistics for management and economics* (3<sup>rd</sup> ed.). North Scituate, MA: Duxbury Press.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- \*Montgomery, D.C., & Runger, G. C. (1994). *Applied statistics and probability for engineers*. New York: John Wiley and Sons, Inc.

Musial, J., III. (1999). Comparing exact tests and asymptotic tests with colorectal cancer variables within the National Health and Nutrition Examination Survey III. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Nanna, M. J. (1997). Robustness and comparative power properties of Hotelling's  $T^2$  versus the rank transformation test using real pre-test/post-test likert scale data. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Nanna, M. J. (2001, in press). Hotelling's  $T^2$  vs the rank transform with real Likert data. *Journal of Modern Applied Statistical Methods*, 1.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation evaluation. *Psychological Methods*, 3, 55-67.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*. London: Unwin Hyman Ltd.

\*Newmark, J. (1988). *Statistics and probability in modern life* (4<sup>th</sup> ed.). New York: Saunders College Publishing.

Posch, M.A., & Sawilowsky, S. (1997). A comparison of exact tests for the analysis of sparse contingency tables. Joint Statistical Meetings, American Statistical Association.

\*Rosenberg, K.M.(1990). *Statistics for behavioral scientists*. Dubuque, IA: Wm. C. Brown Pub.

\*Runyon, R. P. (1977). *Nonparametric statistics: A contemporary approach*. Reading MA: Addison-Wesley Publishing Co.

Sawilowsky, S. S. (1985). Robust and power analysis for the 2x2x2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. Unpublished doctoral dissertation, University of South Florida, Tampa, FL.

Sawilowsky, S. S. (1989). Rank transformation: the bridge is falling down. American Educational Research Association, SIG/Educational Statisticians, San Francisco, CA.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the  $t$  test to departures from population normality. *Psychological Bulletin*, 111, 352-360.

Sawilowsky, S. S., & Brown, M. T. (1991). On using the  $t$  test on ranks as an alternative to the Wilcoxon test. *Perceptual and Motor Skills*, 72, 860-862.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, 14, 255-267.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). REALPOPS.LIB: A PC FORTRAN library of eight real distributions in psychology and education. *Psychometrika*, 55, 729.

Siegel, S. & Castellan, Jr., N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, Inc.

\*Snedecor, G. W. & Cochran, W. G. (1967). *Statistical methods*. Ames, IA: Iowa State University Press.

Sprent, P. (1989). *Applied nonparametric statistical methods*. London: Chapman and Hall.

\*Triola, M. (1995). *Elementary statistics* (6<sup>th</sup> ed.). Reading MA: Addison – Wesley Publishing Company.

\*Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego: Academic Press.

\*Zikmund, W. G. (1991). *Business research methods* (3<sup>rd</sup> ed.). Chicago: The Dryden Press.

<sup>1</sup> Entries with the "\*" refer to the textbook survey results compiled in Table 1, but not cited in this article.

Appendix

Table 3. Samples Randomly Selected from Multimodal Lumpy Data Set (Micceri, 1989)

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
20	11	9	34	10
33	34	14	10	2
4	23	33	38	32
34	37	5	41	4
13	11	8	4	33
6	24	14	26	19
29	5	20	10	11
17	9	18	21	21
39	11	8	13	9
26	33	22	15	31
13	32	11	35	12
9	18	33	43	20
33	27	20	13	33
16	21	7	20	15
36	8	7	13	15

Table 4. Multimodal Lumpy Set (Micceri, 1989).

Score	cum freq	cdf	score	cum freq	cdf
0	5	0.01071	22	269	0.57602
1	13	0.02784	23	279	0.59743
2	21	0.04497	24	282	0.60385
3	24	0.05139	25	287	0.61456
4	32	0.06852	26	297	0.63597
5	38	0.08137	27	306	0.65525
6	41	0.08779	28	309	0.66167
7	50	0.10707	29	319	0.68308
8	62	0.13276	30	325	0.69593
9	80	0.17131	31	336	0.71949
10	91	0.19486	32	351	0.75161
11	114	0.24411	33	364	0.77944
12	136	0.29122	34	379	0.81156
13	160	0.34261	35	389	0.83298
14	180	0.38544	36	401	0.85867
15	195	0.41756	37	418	0.89507
16	213	0.45610	38	428	0.91649
17	225	0.48180	39	434	0.92934
18	234	0.50107	40	445	0.95289
19	244	0.52248	41	454	0.97216
20	254	0.54390	42	460	0.98501
21	261	0.55889	43	467	1.00000

## Adaptive Tests for Ordered Categorical Data

Vance W. Berger  
Biometry Research Group  
National Cancer Institute

Anastasia Ivanova  
Department of Biostatistics  
University of North Carolina

---

Consider testing for independence against stochastic order in an ordered  $2 \times J$  contingency table, under product multinomial sampling. In applications one may wish to exploit prior information concerning the direction of the treatment effect, yet ultimately end up with a testing procedure with good frequentist properties. As such, a reasonable objective may be to simultaneously maximize power at a specified alternative and ensure reasonable power for all other alternatives of interest. For this objective, none of the available testing approaches are completely satisfactory. A new class of admissible adaptive tests is derived. Each test in this class strictly preserves the Type I error rate and strikes a balance between good global power and nearly optimal (envelope) power to detect a specific alternative of most interest. Prior knowledge of the direction of the treatment effect, the level of confidence in this prior information, and possibly the marginal totals might be used to select a specific test from this class.

Key words: Contingency table; exact conditional test; linear rank test; omnibus test; permutation test.

---

### Introduction

When comparing two treatments on the basis of an ordinal endpoint, the data can be summarized as a  $2 \times J$  contingency table. The objective tumor response data, e.g., from 35 ovarian cancer patients treated with cisplatin-based combination chemotherapy and salvage platinum-based therapy (Chiara et al., 1993) are (4,7,2,2) and (1,6,7,6) for patients with treatment-free intervals  $\leq 12$  months and  $> 12$  months, respectively, with categories for 'progressive disease', 'stable disease', 'partial response', and 'complete response'. Combining the two 'non-response' categories, as is common, yields counts  $C_1 = (11,2,2)$  and  $C_2 = (7,7,6)$  in the two groups. For simplicity, the case  $J = 3$  is treated, but with modification the results apply more generally. It is common in practice to dispense with the specification of the alternative hypothesis, and proceed directly to the analysis.

This failure to make the specific alternative hypothesis explicit is unfortunate, because it should serve as the basis for selecting and evaluating the analysis. Linear rank tests, based on assigning numerical scores to the categories, are the most powerful tests to detect point alternatives. If one wishes to test for the superiority of one treatment to another, then stochastic order serves as a reasonable (composite) alternative hypothesis (Cohen and Sackowitz, 1998). Unless the margins satisfy pathological conditions, there is no uniformly most powerful test or monotone likelihood ratio. When testing for stochastic order, nonlinear rank tests, including the Smirnov, improved (Berger and Sackowitz, 1997), convex hull (Berger, Permutt, and Ivanova, 1998; henceforth BPI), and  $COM(L)$  Fisher tests, tend to have better overall power profiles than linear rank tests do.

Berger's (1998) adaptive nonlinear rank test can be generalized to provide an entire class of exact, admissible, adaptive nonlinear rank tests, each of which balances omnibus power for any stochastically ordered alternative against optimal power to detect a specific alternative of greatest interest. The margins may be used to suggest the selection of one particular test from this novel class of tests. The exact conditional powers of some of the aforementioned tests are compared.

---

Vance W. Berger is Mathematical Statistician at the NCI and Adjunct Professor at University of Maryland Baltimore County. E-mail: [vb78c@nih.gov](mailto:vb78c@nih.gov). Anastasia Ivanova is Assistant Professor, Dept. of Biostatistics, School of Public Health., University of North Carolina – Chapel Hill. E-mail: [aivanova@bios.unc.edu](mailto:aivanova@bios.unc.edu).

Notation and Formulation

Consider product multinomial sampling, with  $n_1$  and  $n_2$  (each fixed by the design) patients treated with the control and active treatments, respectively. The vectors of cell probabilities (each summing to one) are  $\pi_1=(\pi_{11},\pi_{12},\pi_{13})$  and  $\pi_2=(\pi_{21},\pi_{22},\pi_{23})$ , respectively, and the corresponding trinomial random vectors are  $C_1 = (C_{11},C_{12},C_{13})$  and  $C_2 = (C_{21},C_{22},C_{23})$ , with  $n_i = C_{i1} + C_{i2} + C_{i3}$ ,  $i = 1, 2$ . The log odds ratios,  $\theta_1$  and  $\theta_2$ , are calculated from  $\pi_1$  and  $\pi_2$  as

$$\theta_1 = \log\{(\pi_{11}\pi_{23})/(\pi_{21}\pi_{13})\} \text{ and}$$

$$\theta_2 = \log\{(\pi_{12}\pi_{23})/(\pi_{22}\pi_{13})\}.$$

Let  $T_j = C_{1j} + C_{2j}$ ,  $j = 1,2,3$ . Conditional on  $T = (T_1,T_2,T_3)$ , the sample space  $\Gamma$  is the set of  $2 \times 3$  contingency tables with nonnegative integer cell counts, and row and column totals  $n = (n_1,n_2)$  and  $T$ , respectively. Given  $T$ ,  $n$ , and  $c = (C_{11},C_{12})$ , the entire  $2 \times 3$  contingency table can be reconstructed as  $C_{13} = n_1 - C_{11} - C_{12}$  and  $C_2 = T - C_1$ . Thus,  $c$  suffices to denote a point of  $\Gamma$ .

Figure 1. The permutation sample space  $\Gamma$  for the data set  $\{(11,2,2);(7,7,6)\}$ , with  $n=(15,20)$  and  $T=(18,9,8)$ .

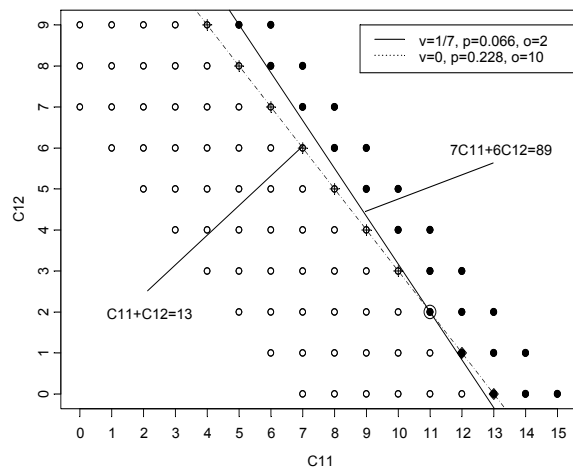


Figure 1 displays  $C_{12}$  plotted against  $C_{11}$  for all 87 tables of  $\Gamma$  for the example,  $\{(11,2,2);(7,7,6)\}$ , with observed table (11,2)

circled. With  $K(T;\theta)=1/\sum_{c \in \Gamma} H(c)\exp[\theta'c]$ ,  $\theta=(\theta_1,\theta_2)$ ,  $\pi=(\pi_1,\pi_2)$ , and  $H(c)=n_1!n_2!/\prod_{i=1}^2 \prod_{j=1}^3 C_{ij}!$ , the density follows the exponential family:

$$P_{\pi}\{c|T\} = P_{\theta}\{c|T\} = K(T;\theta)H(c)\exp[\theta'c]. \quad (2.1)$$

Let  $\Delta_1 = \pi_{11} - \pi_{21}$ , and  $\Delta_2 = (\pi_{11} + \pi_{12}) - (\pi_{21} + \pi_{22}) = \pi_{23} - \pi_{13}$ . If  $\Delta_1 \geq 0$ , and  $\Delta_2 \geq 0$ , at least one strictly, then the active treatment is objectively superior to the control. One may wish to test  $H: \pi_1 = \pi_2$  against the one-sided alternative hypothesis that the active response distribution is stochastically larger than the control response distribution,  $H_A' : \Delta_1 \geq 0, \Delta_2 \geq 0, \pi_1 \neq \pi_2$ . As will be explained, this is not actually possible with a conditional test. By (2.1),  $P_{\pi}\{c|T\}$  depends on  $\pi$  only through  $\theta(\pi)$ , so if  $\theta(\pi) = \theta(\pi^*)$ , then  $c$  offers no information with which to distinguish  $\pi$  from  $\pi^*$ . To be identifiable, then, the hypotheses must be formulated in terms of  $\theta$  (Berger, 1998).

The null hypothesis  $\pi_1 = \pi_2$  is equivalent to  $H: \theta(\pi) = \mathbf{0}$ , but unless  $0 \leq \theta_2 \leq \theta_1$ ,  $\theta(\pi)$  provides insufficient information with which to determine if  $\pi$  satisfies  $H_A'$  because no conditional alternative hypothesis is equivalent to  $H_A$ . Note, e.g., that  $\{(3,3,4)/10;(2,4,4)/10\}$  satisfies  $H_A'$  and  $\{(21,51,328)/400; (7,34,164)/205\}$  does not, yet  $\theta = (\log(3/2),\log(3/4))$  for both. The conditional power to detect  $\pi$  depends on  $\theta(\pi)$  only, so no conditional test that preserves the  $\alpha$ -level whenever  $H_A$  does not hold can be globally powerful whenever it does hold.

However, if  $\pi$  satisfies  $H_A$ , then  $\theta_1(\pi) > 0$ ; and if  $\theta_1 > 0$ , then for any  $\theta_2$  there exists (Berger and Sackowitz, 1997)  $\pi$  satisfying  $H_A$  such that  $\theta(\pi) = (\theta_1,\theta_2)$ . As such,  $\theta_1$  is the key parameter; the active treatment is superior on  $\Omega_A = \{\theta | \theta_1 >$

0}, no different on  $\Omega_0 = \{\theta | \theta_1 = 0\}$ , and inferior on  $\Omega_C = \{\theta | \theta_1 < 0\}$ . It is reasonable, then, to test  $H$  against  $H_A : \theta_1 > 0$ . The large unconditional indifference region, where neither group stochastically dominates the other, has, by conditioning, been absorbed into  $\Omega_0 \cup \Omega_A \cup \Omega_C$ .

Let  $\delta(\theta) = 1 - \theta_2/\theta_1$  be the *direction* of the effect. As  $\theta_1$  increases in both  $\Delta_1$  and  $\Delta_2$ , while  $\theta_2$  ( $\theta_1 - \theta_2$ ) increases in  $\Delta_2$  ( $\Delta_1$ ), and decreases in  $\Delta_1$  ( $\Delta_2$ ), the superiority of the active treatment to the control is due primarily to a shift from the middle to the best outcome ( $\Delta_2 > \Delta_1$ ) if  $\delta(\theta)$  is small, or from the worst to the middle outcome ( $\Delta_1 > \Delta_2$ ) if  $\delta(\theta)$  is large. Let  $\Omega_v = \{\theta | \theta_1 > 0, \delta(\theta) = v\}$ . As  $\delta(\theta)$  is generally unknown *a priori*, omnibus tests that are sensitive to departures from  $H_0$  in each direction of  $\Omega_A = \cup_{v \in \mathfrak{R}^1} \Omega_v$  are preferred to tests that lack this desirable property.

If the  $\phi$  rejection region  $R_\alpha(\phi)$  contains  $D[\Gamma]$ , the set of directed extreme points of  $\Gamma$  (BPI, 1998), then  $\phi$  is omnibus. The challenge is to exploit prior information about  $\delta(\theta)$  to construct omnibus tests with especially good power in one preferred direction,  $\Omega_v$ . For reasons articulated by Berger (2000) and Berger et al. (2002), we consider only exact conditional tests in this formulation.

A New Look at Linear Rank Tests

Linear rank tests are based on numerical scores  $(v_1, v_2, v_3)$ ,  $v_1 < v_3$ , assigned to the three outcome levels. With  $v = (v_2 - v_1)/(v_3 - v_1)$ ,  $\phi_v$  uses test statistic  $z_v(\mathbf{c}) = C_{11} + (1 - v)C_{12}$ . New notation allows for greater insight into linear rank tests. Let  $M_v(\mathbf{c}) = \{\mathbf{c}^* \in \Gamma | z_v(\mathbf{c}^*) \geq z_v(\mathbf{c})\}$  be the  $\phi_v$  extreme region of  $\mathbf{c}$ , with boundary  $B_v(\mathbf{c})$  and p-value  $p_v(\mathbf{c}) = P_{\mathbf{0}}\{M_v(\mathbf{c})|T\}$ . The level set (Frick, 2000, p. 719) of  $z_v(\mathbf{c})$  is  $B_v(\mathbf{c}) \cap \Gamma$ , with  $o_v(\mathbf{c})$  its order, or the number of points of  $B_v(\mathbf{c}) \cap \Gamma$ . If  $\mathbf{c} =$

$(C_{11}, C_{12}) \in \Gamma$  and  $\mathbf{c}^* = (C_{11}^*, C_{12}^*) \in \Gamma - \mathbf{c}$ , then  $z_v(\mathbf{c}^*) = z_v(\mathbf{c})$  if and only if  $v = 1 - (C_{11} - C_{11}^*)/(C_{12}^* - C_{12})$ , say  $v = v_{\mathbf{c}, \mathbf{c}^*}$  (vector valued for  $J > 3$ ). Let  $V(\mathbf{c}) = \{v_1(\mathbf{c}), v_2(\mathbf{c}), \dots, v_{K_c}(\mathbf{c})\}$  be the ordered set  $\{v_{\mathbf{c}, \mathbf{c}^*} | |v_{\mathbf{c}, \mathbf{c}^*}| < \infty, \mathbf{c}^* \in \Gamma - \mathbf{c}\}$ , and let  $v_0(\mathbf{c}) = -\infty$  and  $v_{K_c+1}(\mathbf{c}) = \infty$ . For finite  $v$ ,  $o_v(\mathbf{c}) > 1$  if and only if  $v \in V(\mathbf{c})$ .

Let  $\varepsilon(\mathbf{c}) = \min_k [v_{k+1}(\mathbf{c}) - v_k(\mathbf{c})]/2$ ,  $z_v^\perp(\mathbf{c}) = C_{12} + (v - 1)C_{11}$ ,  $B_v^+(\mathbf{c}) = \{\mathbf{c}^* \in B_v(\mathbf{c}) \cap \Gamma | z_v^\perp(\mathbf{c}^*) > z_v^\perp(\mathbf{c})\}$ ,  $B_v^-(\mathbf{c}) = \{\mathbf{c}^* \in B_v(\mathbf{c}) \cap \Gamma | z_v^\perp(\mathbf{c}^*) < z_v^\perp(\mathbf{c})\}$ ,  $v^*(\mathbf{c}) = \{v^* | p_{v^*}(\mathbf{c}) \leq p_v(\mathbf{c}) \text{ for all } v^*\}$ .

By Lemma 1 (in the Appendix),  $v^*(\mathbf{c})$  consists of the scores that minimize not just  $p_v(\mathbf{c})$  but also  $p_{\min(v)}(\mathbf{c}) = \min(\lim_{u \downarrow v} p_u(\mathbf{c}), \lim_{u \uparrow v} p_u(\mathbf{c})) = p_v(\mathbf{c}) - \max(P_0\{B_v^-(\mathbf{c})\}, P_0\{B_v^+(\mathbf{c})\})$ . Hence,  $p_{\min(v)}(\mathbf{c})$ , which also equals  $\min\{p_{v-\varepsilon}(\mathbf{c}), p_{v+\varepsilon}(\mathbf{c})\}$ , is a true p-value. As  $\Gamma$  has finitely many subsets, there can be only a finite number of values for  $p_v(\mathbf{c})$ , so the minimum p-value is attained, and  $v^*(\mathbf{c}) \neq \emptyset$ . If  $v \in V(\mathbf{c})$ , then  $o_v(\mathbf{c}) > 1$ ,  $B_v^-(\mathbf{c}) \cup B_v^+(\mathbf{c}) \neq \emptyset$ ,  $p_{\min(v)}(\mathbf{c}) < p_v(\mathbf{c})$ , and  $v \notin v^*(\mathbf{c})$ . Hence,  $v^*(\mathbf{c}) \cap V(\mathbf{c}) = \emptyset$ , and, by Lemma 1,  $v^*(\mathbf{c})$  consists of one or more open intervals of the form  $(v_k(\mathbf{c}), v_{k+1}(\mathbf{c}))$ . For  $\{(11,2,2);(7,7,6)\}$ ,  $\mathbf{c} = (11,2)$ ,  $K_c = 42$ ,  $\varepsilon(11,2) = 1/84$ , and  $V(\mathbf{c}) = \{-6, -5, -4, -3, -5/2, -2, -5/3, -3/2, -4/3, -5/4, -6/5, -1, -5/6, -4/5, -3/4, -2/3, -3/5, -4/7, -1/2, -3/7, -2/5, -1/3, -2/7, -1/4, -1/5, -1/6, -1/7, 0, 1/7, 1/6, 1/5, 1/4, 2/7, 1/3, 2/5, 1/2, 2/3, 1, 3/2, 2, 5/2, 3, 4, 5, 6\}$ .

Figure 1 shows  $M_{1/7}(11,2)$  by dark dots and  $M_0(11,2) - M_{1/7}(11,2)$  by crosses. Because (11,2) minimizes  $z_{1/7}^\perp(11,2) = 7C_{12} - 6C_{11}$  over  $B_{1/7}(11,2) \cap \Gamma$  (Table 1),  $B_{1/7}^-(11,2) = \emptyset$  and  $p_{1/7}(11,2) = \lim_{u \uparrow 1/7} p_u(11,2) = 0.066$ . Also  $p_v(11,2) = 0.020$  for  $v \in (1.0, 1.5) = v^*(11,2)$ . If  $v$

$\in V(11,2)$ , then  $P_0\{B_v^-\} \leq P_0\{B_v^+\}$  for  $v > 1.5$ , and  $P_0\{B_v^-\} \geq P_0\{B_v^+\}$  for  $v < 1.0$ . The optimality of most powerful (MP) test  $\phi_{\delta}(\theta)$  to detect  $I\theta$ , for  $I > 0$  (BPI, 1998), is offset by its

potentially poor power on  $\Omega_A - \Omega_{\delta}(\theta)$ . In fact,  $D[\Gamma]$  may not be contained in the  $\phi_v$  critical region  $R_{\alpha}(\phi_v)$  for any  $v$ , so for

Table 1. All possible linear rank tests with scores  $(0,v,1)$ , with middle score  $v \in [0,2]$ , for the data set  $\{(11,2,2);(7,7,6)\}$ , along with the number of points in its level set, the endpoints and null probabilities of each segment of its level set, and various p-values. (null probabilities of various extreme regions).

$v$	$\alpha_v(11,2)$	Endpoints of: $B_v^+$ $B_v^-$	$p_v$ (minimum is underlined)	$p_v^-$	$p_v^+$	$P_0\{B_v^+\}$	$P_0\{B_v^-\}$	$p_{v,\infty}$	$M_v - M_{v,\infty}$
$v \in (-1/7,0)$	1		0.2262	0.2262	0.2262			0.2262	
$v = 0$	10	(4,9)   (12,1) -(10,3)   -(13,0)	0.2277	0.2262	<u>0.0661</u>	0.1615	0.0015	0.0726	(7,6)- (10,3)
$v \in (0,1/7)$	1		0.0661	0.0661	0.0661			0.0661	
$v = 1/7$	2	(5,9)	0.0661	0.0661	<u>0.0661</u>	$2.1 \cdot 10^{-5}$		0.0661	
$v \in (1/7,1/6)$	1		0.0661	0.0661	0.0661			0.0661	
$v = 1/6$	2	(6,8)	0.0661	0.0661	<u>0.0657</u>	0.0004		0.0661	
$v \in (1/6,1/5)$	1		0.0657	0.0657	0.0657			0.0657	
$v = 1/5$	2	(7,7)	0.0657	0.0657	<u>0.0629</u>	0.0028		0.0657	
$v \in (1/5,1/4)$	1		0.0629	0.0629	0.0629			0.0629	
$v = 1/4$	2	(8,6)	0.0629	0.0629	<u>0.0538</u>	0.0091		0.0629	
$v \in (1/4,2/7)$	1		0.0538	0.0538	0.0538			0.0538	
$v = 2/7$	2	(6,9)	0.0538	0.0538	<u>0.0538</u>	$5.7 \cdot 10^{-6}$		0.0538	
$v \in (2/7,1/3)$	1		0.0538	0.0538	0.0538			0.0538	
$v = 1/3$	3	(7,8) -(9,5)	0.0538	0.0538	<u>0.0387</u>	0.0152		0.0387	(9,5)
$v \in (1/3,2/5)$	1		0.0387	0.0387	0.0387			0.0387	
$v = 2/5$	2	(8,7)	0.0387	0.0387	<u>0.0382</u>	0.0005		0.0387	
$v \in (2/5,1/2)$	1		0.0382	0.0382	0.0382			0.0382	
$v = 1/2$	4	(9,6)   (12,0) -(10,4)	0.0385	0.0382	<u>0.0237</u>	0.0148	0.0003	0.0249	(10,4)
$v \in (1/2,2/3)$	1		0.0237	0.0237	0.0237			0.0237	
$v = 2/3$	2	(10,5)	0.0237	0.0237	<u>0.0220</u>	0.0017		0.0237	
$v \in (2/3,1)$	1		0.0220	0.0220	0.0220			0.0220	
$v = 1$	5	(11,4)   (11,1) -(11,3)   -(11,0)	0.0276	0.0220	<b><u>0.0198</u></b>	0.0078	0.0056	0.0276	
$v \in (1,3/2)$	1		<b><u>0.0198</u></b>	<b><u>0.0198</u></b>	<b><u>0.0198</u></b>			0.0198	
$v = 3/2$	2	(10,0)	0.0205	<b><u>0.0198</u></b>	0.0205		0.0008	0.0205	
$v \in (3/2,2)$	1		0.0205	0.0205	0.0205			0.0205	
$v = 2$	4	(12,3)   (10,1) -(9,0)	0.0294	<u>0.0205</u>	0.0289	0.0005	0.0089	0.0294	
$v \in (2,5/2)$	1		0.0289	0.0289	0.0289			0.0289	

Note that all the values are calculated at the outcome  $(11,2)$ ;  $p_{v,\infty}$  and  $M_{v,\infty}$  are the p-value and extreme region, respectively, of the adaptive test based on  $v$  and  $\tau = \infty$ .



each  $v$  there will exist  $\theta \in \Omega_A$  for which the power of  $\varphi_v$  to detect  $l\theta$  tends to zero as  $l$  gets large (BPI, 1998). Podgor, Gastwirth, and Mehta (1996) proposed the maximin efficiency robust test (MERT) in hopes of providing better power than linear rank tests. Ironically, the MERT is itself a linear rank test; its rejection region may also fail to contain  $D[\Gamma]$ , leading to poor power on parts of  $\Omega_A$  and no power in the limit in some directions. Berger and Ivanova (2002) showed that at certain  $\alpha$ -levels the most stringent linear rank test is  $\varphi_{v_S}$ , where  $v_S$  is such that the two points of  $D[\Gamma]$  that are furthest (in Euclidean distance) from each other are equated by  $z_{v_S}(c)$ . For  $\{(11,2,2),(7,7,6)\}$ , this gives  $v_S = 0$ , because  $\Gamma$  has two directed extreme points,  $D[\Gamma]=\{(15,0);(6,9)\}$ , and  $z_0(15,0) = 15+(1-0)(0) = 15 = 6+(1-0)(9) = z_0(6,9)$ .

#### Nonlinear Rank Tests

By allowing the boundary of  $R_\alpha(\varphi)$  to curve, nonlinear rank tests often require smaller  $\alpha$ -levels to ensure that  $D[\Gamma] \subset R_\alpha(\varphi)$  than linear rank tests would. However, this is not always the case. Berger and Ivanova (2002) provide an example in which the proportional odds and proportional hazards tests (McCullagh, 1980) are not nonlinear enough to be omnibus at reasonable  $\alpha$ -levels. The Smirnov test,  $\varphi_S$ , uses as the test statistic the largest of three quantities, 0,  $D_1 = C_{11}/n_1 - C_{21}/n_2$ , and  $D_2 = (C_{11} + C_{12})/n_1 - (C_{21} + C_{22})/n_2$ . Among tests routinely available in standard statistical software packages ( $\varphi_S$  is a standard feature of StatXact),  $\varphi_S$  minimizes the  $\alpha$ -level required for its rejection region to contain  $D[\Gamma]$ . However,  $\varphi_S$  is not generally admissible (Berger, 1998).

Permutt and Berger (2000) and Ivanova and Berger (2001) each proposed refinements of  $\varphi_S$  that break its ties. Although such refinements are necessarily uniformly more powerful than  $\varphi_S$  (Rohmel and Mansmann, 1999, p. 158), the term

“improvement of  $\varphi$ ” is reserved for a test whose exact (possibly randomized) version is uniformly more powerful than the exact (possibly randomized) version of  $\varphi$ . By this definition, refinements are rarely improvements. Berger and Sackrowitz (1997) developed methodology for constructing improvements of a given inadmissible test. In fact, by improving the “ignore-the-data” test,  $\varphi_{ITD}(c) = \alpha$  for all  $c \in \Gamma$ , Berger and Sackrowitz (1997) constructed the first known test for this problem that is simultaneously admissible and unbiased. However, rejection regions at different  $\alpha$ -levels need not be nested, so these improved tests may not yield unambiguous p-values, and thus are of somewhat limited value.

Berger (1998) established the one-to-one correspondence between the class of convex hull type tests and the minimal complete class of admissible tests. The convex hull test (BPI, 1998),  $\varphi_{CH}$ , is the simplest member of this convex hull class, and is qualitatively similar to the improvements of both  $\varphi_S$  and  $\varphi_{ITD}$ , while minimizing, among all families of tests, the  $\alpha$ -level required for its rejection region to contain  $D[\Gamma]$ .

In addition,  $\varphi_{CH}$  is based on a test statistic, so rejection regions at different  $\alpha$ -levels are nested, and p-values are provided. As such,  $\varphi_{CH}$  is about as good a test as there is for testing  $H$  against  $H_A$ , which is about as close as one can get to testing  $H$  against  $H'_A$  when dealing with  $\theta$  instead of  $\pi$ . Specifically, admissible (unbiased) tests of  $H$  against  $H_A$  are conditionally admissible (unbiased) as tests of  $H$  against  $H'_A$  (Berger and Sackrowitz, 1997). However,  $\theta(\pi)$  is a nonlinear function, and maps small corners of  $\pi$ -space (neighborhoods of structural zeros) into large regions of  $\theta$ -space. By giving each direction  $\delta(\theta)$  equal consideration,  $\varphi_{CH}$  accommodates these small corners as much as it does the large regions of  $\pi$ -space that are of greatest unconditional interest. As such,  $\varphi_{CH}$  may not be ideal when viewed unconditionally. Cohen and Sackrowitz (1998) proposed another member of the convex hull class, called the *COM(L)* Fisher test, or

$\varphi_{COM(L)}$ , based on repeatedly adding to the critical region those directed extreme points of the current acceptance region that are least likely under  $H_0$ . Because the test statistics of  $\varphi_{COM(L)}$  and  $\varphi_{CH}$  are defined not algebraically but relationally, by the relative position of  $c$  within  $\Gamma$ , the rejection regions need to be constructed recursively. This feature is a barrier to their use.

Adaptive Tests

Gross (1981, Section 5) suggested that an "analysis based on ... data-dependent scores may yield procedures that compare favorably to fixed-score procedures ...". Distinct from another definition used, e.g., by Rukhin and Mak (1992), Hogg (1974, p. 917) and Edgington (1995, pp. 371-373) defined adaptive tests as tests with data-based test statistics. This allows  $\Gamma$  to be partitioned into regions sharing a common test statistic. Because the region need not be even nearly ancillary, conditioning on the region (as suggested by Donegani, 1991, and Good, 1994, p. 122) may entail a loss of power. Comparing the value of the test statistics across regions avoids this loss of power. The intuitive objection to "comparing apples to oranges" notwithstanding, such an approach is "good" or "bad" only to the extent to which it produces a "good" or "bad" test. This approach results in tests with excellent power properties. In fact, Gastwirth (1985) stated that "when the MERT for a particular problem has a low  $r^2$ , adaptive procedures are needed".

Without knowing  $\theta$  *a priori*, it is unclear where to maximize the power. One could estimate  $\delta(\theta)$  from  $c$ , say as  $\hat{\delta}_p(c)$ , perhaps using maximum likelihood, and use the MP test  $\varphi_{\hat{\delta}_p}$ . The p-value of  $\varphi_{\hat{\delta}_p}$  evaluated at observed outcome  $c$ ,  $p_{\hat{\delta}_p}(c)$ , is stochastically too small to serve as a valid p-value, but  $p_{\hat{\delta}_p}(c)$  can be used as a *test statistic*, to be compared to its null distribution (Rohmel and Mansmann, 1999, p. 165). Variation in  $c$  is reflected in  $p_{\hat{\delta}_p}(c)$  through *both* the argument and the subscript. Using either  $p_{\hat{\delta}_p}(c)$  or  $z_{\hat{\delta}_p}(c)$ , suitably normalized, as a test statistic, any

estimator  $\hat{\delta}_p(c)$  of  $\delta(\theta)$  induces an adaptive test, with regions  $\Gamma_v = \hat{\delta}^{-1}(v) = \{c \in \Gamma \mid \hat{\delta}_p(c) = v\}$ . If the regions are  $\Gamma_0 = \{c \in \Gamma \mid C_{12} > n_1 T_2 / (n_1 + n_2)\}$ ,  $\Gamma_1 = \Gamma - \Gamma_0$ , and  $\Gamma_v = \emptyset$  for  $v \notin \{0,1\}$ , and the  $\varphi_v$  test statistic  $z_v(c)$  is used on  $\Gamma_v$ , with  $C_{11} + C_{12}$  ( $v = 0$ ) and  $C_{11}$  ( $v = 1$ ) normalized to  $D_2$  and  $D_1$ , respectively, to facilitate the comparison of points from  $\Gamma_1$  ( $D_1 > D_2$ ) to those from  $\Gamma_0$  ( $D_2 \geq D_1$ ), then  $\varphi_S$  results. Similar binary adaptive tests might define  $\Gamma_0$  and  $\Gamma_1$  by whichever of  $\varphi_0$  and  $\varphi_1$  yields a smaller p-value or a larger  $\chi^2$ .

Berger (1998) proposed judging outcome  $c$  by how small a p-value it can yield with an MP test; that is,  $\varphi_A$  uses  $p_{v^*}(c) = \min_{-\infty \leq v \leq \infty} p_v(c)$  as the test statistic. This is a continuous version of the adaptive test based on  $\min(p_0(c), p_1(c))$ , and estimates  $\delta(\theta)$  non-uniquely as  $\hat{\delta}_c = v$  for any value  $v \in v^*(c)$ . The induced regions are  $\Gamma_v = \{c \in \Gamma \mid v \in v^*(c)\}$ . The  $\varphi_A$  critical region is  $R_\alpha(\varphi_A) = \cup_{v \in R^1} R_{\alpha^*}(v)(\varphi_v)$  for some set of  $\alpha^*(v) < \alpha$ , so  $\varphi_A$  is intuitively similar to union-intersection tests (Roy, 1953; Marden, 1991). Despite being constructed non-recursively,  $\varphi_A$  is a convex hull type test (Berger, 1998); hence  $\varphi_A$  is always admissible. Also,  $\varphi_A$  tends to be omnibus, as  $D[\Gamma] \subset R_\alpha(\varphi_A)$  for reasonable  $\alpha$ -levels.

Accommodating a Favored Alternative

Suppose that one believes *a priori* that  $\delta(\theta) = \delta_p$ . Let  $\tau \geq 0$  be a measure of the strength in the belief that  $\delta(\theta) = \delta_p$ . The dual objectives are ensuring nearly MP power on  $\Omega_{\delta_p}$  and reasonable power on  $\Omega_A - \Omega_{\delta_p}$ , with relative importance dictated by  $\tau$ . One might use  $\varphi_{\delta_p}$  (which is MP on  $\Omega_{\delta_p}$ ) for large  $\tau$ , or  $\varphi_A$  (which is a good omnibus test) for small  $\tau$ , but none of the aforementioned test suffices for intermediate values of  $\tau$ . Linear

combinations such as  $(\tau \varphi_{\delta_p} + \varphi_A)/(\tau + 1)$  would not suffice either, because they have large randomization regions and small critical regions, consisting only of the intersection  $R_\alpha(\varphi_{\delta_p}) \cap R_\alpha(\varphi_A)$ . Of course, these inadmissible tests could be improved to admissibility, but then the procedure would be complicated, and p-values may not be defined. There is another approach to bridge the gap between  $\varphi_{\delta_p}$  and  $\varphi_A$ . Specifically, start with  $\varphi_A$ , but penalize those  $\mathbf{c}$  whose minimizing MP p-value is obtained by  $v$  far from  $\delta_p$ . To this end, let  $\varphi_{\delta_p, \tau, \alpha}$  (or  $\varphi_{\delta_p, \tau}$ ) be the level- $\alpha$  adaptive test based on the test statistic

$$A(\delta_p, \tau, \mathbf{c}) = \min_{-\infty \leq v \leq \infty} [\rho_{\min(v)}(\mathbf{c})(1 + |\delta_p - v|)^\tau].$$

Let  $v_{[\delta_p, \tau]}(\mathbf{c}) = \{v \mid p_{\min(v)}(\mathbf{c})(1 + |\delta_p - v|)^\tau = A(\delta_p, \tau, \mathbf{c})\}$ . Clearly,  $\varphi_{\delta_p, 0} = \varphi_A$  for any  $\delta_p$  and  $p_{\min(v)}(\mathbf{c})(1 + |\delta_p - v|)^\tau \leq 1$  if  $v \in v_{[\delta_p, \tau]}(\mathbf{c})$ . Lemmas 2-4 confine  $v_{[\delta_p, \tau]}(\mathbf{c})$  to a finite subset of an interval that shrinks, as  $\tau$  gets large, to  $\{\delta_p\}$ . By Lemma 4,  $\varphi_{\delta_p, \infty}$  induces the same ordering on  $\Gamma$  as  $\varphi_{\delta_p}$  does, thereby optimizing power on  $\Omega_{\delta_p}$ . Yet because the  $\varphi_{\delta_p, \infty}$  test statistic is  $p_{\min(\delta_p)}(\mathbf{c})$ , and not necessarily  $p_{\delta_p}(\mathbf{c})$ ,  $\varphi_{\delta_p, \infty}$  is a refinement of  $\varphi_{\delta_p}$ , and  $p_{\min(v)}(\mathbf{c}) \leq p_{v, \infty}(\mathbf{c}) \leq p_v(\mathbf{c})$  for all  $v$  and  $\mathbf{c}$ . From Table 1, e.g.,  $p_{0.5}(11, 2) = 0.0385$ , but  $p_{0.5, \infty}(11, 2) = 0.0385 - P_0\{(10, 4) | T\} = 0.0249$ . Each test in the class of adaptive tests is admissible.

*Theorem 1.* For any triple  $\delta_p \in \mathfrak{R}^1$ ,  $\tau \geq 0$ , and  $\alpha \in [0, 1]$ ,  $\varphi_{\delta_p, \tau, \alpha}$  is admissible. Graubard and Korn (1987) suggested that without a reason to use a different  $\delta_p$ ,  $\varphi_{0.5}$  should be used. The desire to focus power on the "central" direction,  $\Omega_{0.5}$ , is understandable, but the use of linear rank tests in general (BPI, 1998; Berger and Ivanova, 2002), and  $\varphi_{0.5}$  in particular (Ivanova and Berger, 2001), have been criticized. Now  $\varphi_{0.5, \tau}$  offers good central power without sacrificing global power

(unless  $\tau = \infty$ ). But even if  $\tau = \infty$ ,  $\varphi_{0.5, \infty}$  is still more powerful than, and hence preferable to  $\varphi_{0.5}$ .

#### Margin-Based Selection of $\delta_p$ and $\tau$

Recall that  $v_S$  can be determined from the margins ( $\mathbf{n}$  and  $\mathbf{T}$ , summarized by  $\Gamma$ ). In some cases, it may be reasonable to use  $v_S$  as  $\delta_p$ . In others, it may be reasonable to use the margins to find the largest  $\tau$  that allows  $R_\alpha(\varphi_{\delta_p, \tau, \alpha})$  to contain  $D[\Gamma]$ . Unless  $|\delta_p - v_S|$  is small, the larger  $\tau$  is, the less  $\varphi_{\delta_p, \tau}$  focuses on omnibus power. Hence, the  $\alpha$ -level required for  $R_\alpha(\varphi_{\delta_p, \tau, \alpha})$  to contain  $D[\Gamma]$  tends to increase in  $\tau$ . If a range of  $\alpha$ -levels would be considered, say  $0.01 \leq \alpha \leq 0.1$ , then use the smallest  $\alpha$ -level in selecting  $\tau$ . Restricting attention to the integer values of  $\tau$ , and using  $\delta_p = 0.5$ , note that for  $\{(11, 2, 2), (7, 7, 6)\}$ ,  $D[\Gamma] = \{(6, 9); (15, 0)\}$  is contained by  $R_{0.01}(\varphi_{0.5, 18})$ ,  $R_{0.025}(\varphi_{0.5, 20})$ ,  $R_{0.05}(\varphi_{0.5, 22})$ , and  $R_{0.1}(\varphi_{0.5, 24})$ ; but none of  $R_{0.01}(\varphi_{0.5, 19})$ ,  $R_{0.025}(\varphi_{0.5, 21})$ ,  $R_{0.05}(\varphi_{0.5, 23})$ , or  $R_{0.1}(\varphi_{0.5, 25})$  contain  $(6, 9)$ . Consequently,  $\varphi_{0.5, 18}$  would be used by this approach.

#### Comparisons of Tests

The exact conditional power of the one-sided nonrandomized versions of  $\varphi_{0.0}$ ,  $\varphi_{0.5}$ ,  $\varphi_{1.0}$ ,  $\varphi_S$ ,  $\varphi_{CH}$ ,  $\varphi_{COM(L)}$ , and some adaptive tests, at  $\alpha \leq 0.05$ , are compared considering all 87  $2 \times 3$  tables with row and column margins as in the example,  $T = (18, 9, 8)$ ,  $\mathbf{n} = (15, 20)$ . Figure 2 illustrates extreme regions. The exact conditional power of  $\varphi$  to detect  $\theta$  is calculated as  $P_\theta\{R_{0.05}(\varphi) | T\}$ . Here  $4 \times 7 = 28$  alternatives, with  $\theta_1 \in \{0.5, 1.0, 1.5, 2.0\}$  and  $\theta_2 = \{-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5\}$ , are considered, along with the null case,  $\theta_1 = \theta_2 = 0$ . Bold entries represent the best power, for given  $\theta$ , among the six targeted tests in columns 4-9 and among five omnibus tests in columns 10-13. Because the linear rank tests  $\varphi_{0.0}$  ( $\alpha = 0.005$ ),  $\varphi_{0.5}$  ( $\alpha = 0.038$ ), and  $\varphi_{1.0}$  ( $\alpha = 0.028$ ) are excessively conservative, per the top

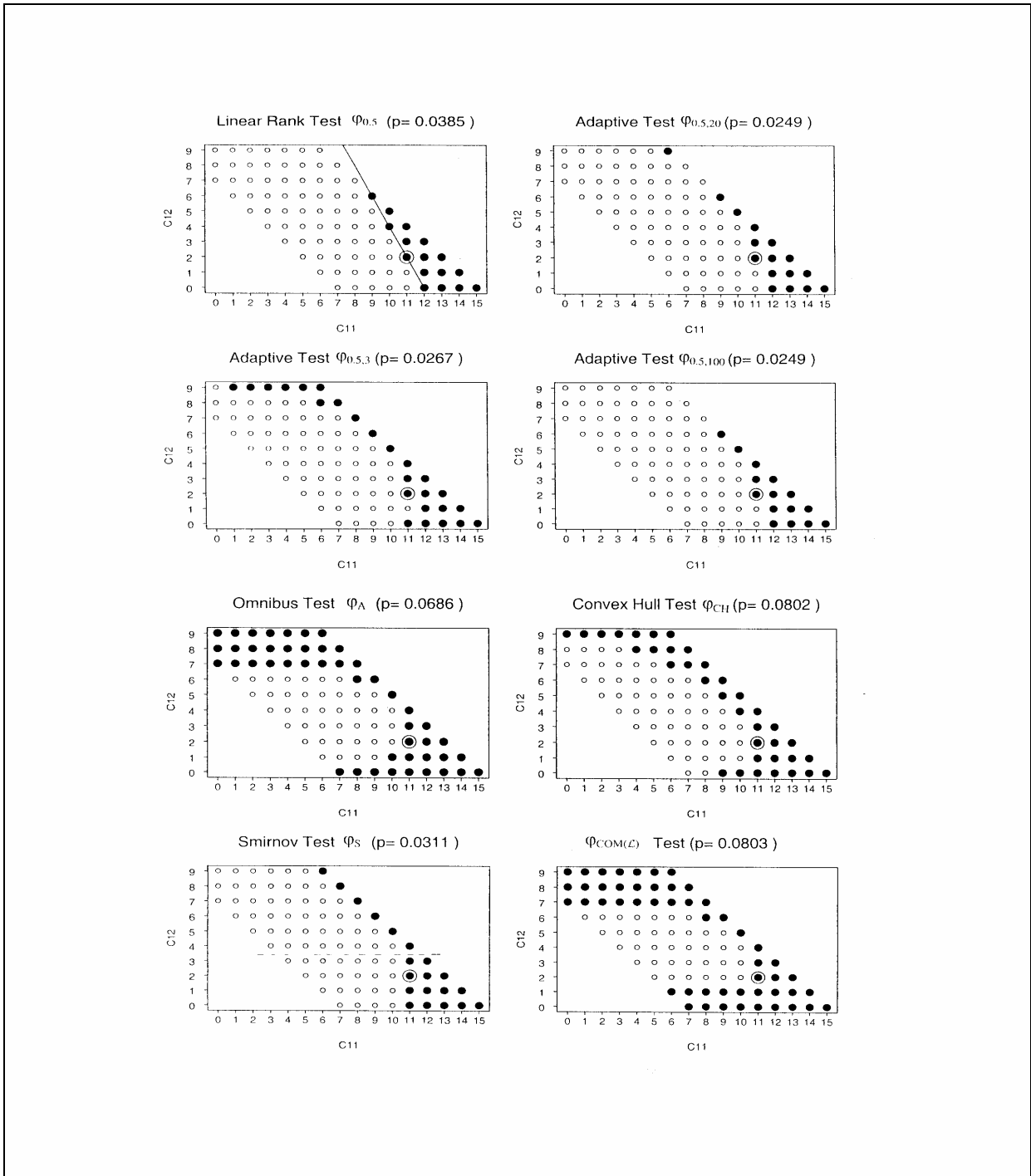


Figure 2. Extreme regions and p-values for  $\{(11,2,2);(7,7,6)\}$  and several tests including the linear rank test with equally-spaced scores  $\varphi_{0,5}$ , the adaptive tests with similar direction but varying second parameter  $\varphi_{0.5,3}$ ,  $\varphi_{0.5,20}$ ,  $\varphi_{0.5,100}$ , the omnibus adaptive test  $\varphi_A$ , the Smirnov test  $\varphi_S$ , the convex hull test  $\varphi_{CH}$ , and the  $\varphi_{COM(L)}$  test.

Table 2. Exact conditional power of the conservative (nonrandomized) versions of linear rank tests ( $\varphi_0, \varphi_1, \varphi_{0.5}$ ), adaptive tests ( $\varphi_{0,100}, \varphi_{1,100}, \varphi_{0.5,100}, \varphi_{0.5,1}$ ), omnibus adaptive test  $\varphi_A$ , the  $\varphi_{COM(L)}$  test, Smirnov test  $\varphi_S$ , and convex hull test  $\varphi_{CH}$ , with  $\alpha \leq 0.05$ , and table margins  $T=(18,9,8)$ ,  $n=(15,20)$ . Bold entries represent the best power among the tests in each block (narrow and omnibus) for each given  $\theta$ .

$\delta(\theta)$	$\theta$	$\varphi_0$	$\varphi_{0,100}$	$\varphi_{0.5}$	$\varphi_{0.5,100}$	$\varphi_1$	$\varphi_{1,100}$	$\varphi_{0.5,1}$	$\varphi_A$	$\varphi_{COM(L)}$	$\varphi_S$	$\varphi_{CH}$
	0.0 0.0	0.005	0.040	0.038	0.044	0.028	0.039	0.046	0.047	0.050	0.031	0.035
-2.000	0.5 1.5	0.054	<b>0.232</b>	0.046	0.063	0.006	0.015	0.258	<b>0.375</b>	0.316	0.058	0.255
-1.000	0.5 1.0	0.038	<b>0.163</b>	0.071	0.080	0.021	0.032	0.150	<b>0.198</b>	0.152	0.053	0.145
-0.500	1.0 1.5	0.107	<b>0.325</b>	0.151	0.174	0.039	0.067	0.290	<b>0.332</b>	0.244	0.131	0.285
0.000	0.5 0.5	0.025	<b>0.120</b>	0.103	0.110	0.057	0.070	<b>0.109</b>	0.108	0.090	0.073	0.093
0.000	1.0 1.0	0.079	<b>0.264</b>	0.212	0.223	0.099	0.126	<b>0.219</b>	0.215	0.169	0.151	0.200
0.000	1.5 1.5	0.184	<b>0.447</b>	0.352	0.371	0.149	0.208	<b>0.366</b>	0.361	0.292	0.270	0.349
0.250	2.0 1.5	0.280	0.606	0.603	<b>0.615</b>	0.370	0.445	<b>0.524</b>	0.491	0.460	0.485	0.489
0.333	1.5 1.0	0.143	0.417	0.442	<b>0.455</b>	0.288	0.328	<b>0.379</b>	0.333	0.310	0.346	0.330
0.500	1.0 0.5	0.055	0.231	0.274	<b>0.291</b>	0.200	0.225	<b>0.244</b>	0.196	0.189	0.223	0.193
0.500	2.0 1.0	0.231	0.593	0.689	<b>0.704</b>	0.560	0.597	<b>0.615</b>	0.543	0.537	0.605	0.542
0.667	1.5 0.5	0.109	0.390	0.521	<b>0.550</b>	0.454	0.481	<b>0.483</b>	0.391	0.395	0.475	0.390
0.750	2.0 0.5	0.188	0.560	0.754	<b>0.785</b>	0.723	0.741	<b>0.738</b>	0.634	0.640	0.735	0.634
1.000	0.5 0.0	0.015	0.096	0.137	<b>0.157</b>	0.121	0.147	<b>0.140</b>	0.104	0.116	0.127	0.100
1.000	1.0 0.0	0.038	0.201	0.333	<b>0.378</b>	0.332	0.368	<b>0.347</b>	0.258	0.283	0.339	0.257
1.000	1.5 0.0	0.082	0.349	0.585	<b>0.646</b>	0.612	0.642	<b>0.621</b>	0.499	0.521	0.617	0.499
1.000	2.0 0.0	0.153	0.514	0.799	0.851	0.836	<b>0.852</b>	<b>0.841</b>	0.736	0.748	0.839	0.736
1.250	2.0 -0.5	0.126	0.467	0.830	0.896	0.906	<b>0.924</b>	<b>0.908</b>	0.828	0.844	0.906	0.828
1.333	1.5 -0.5	0.062	0.302	0.634	0.729	0.736	<b>0.779</b>	<b>0.744</b>	0.628	0.665	0.737	0.628
1.500	1.0 -0.5	0.026	0.167	0.384	0.472	0.471	<b>0.536</b>	<b>0.483</b>	0.377	0.432	0.472	0.375
1.500	2.0 -1.0	0.106	0.429	0.854	0.920	0.944	<b>0.963</b>	<b>0.948</b>	0.899	0.915	0.944	0.899
1.667	1.5 -1.0	0.048	0.262	0.671	0.784	0.822	<b>0.876</b>	<b>0.834</b>	0.752	0.795	0.823	0.750
1.750	2.0 -1.5	0.093	0.401	0.874	0.930	0.965	<b>0.983</b>	<b>0.970</b>	0.945	0.958	0.965	0.945
2.000	0.5 -0.5	0.010	0.077	0.171	0.221	0.212	<b>0.273</b>	<b>0.227</b>	0.167	0.217	0.214	0.163
2.000	1.0 -1.0	0.018	0.136	0.426	0.552	0.593	<b>0.692</b>	<b>0.617</b>	0.524	0.604	0.593	0.520
2.000	1.5 -1.5	0.038	0.231	0.703	0.811	0.877	<b>0.934</b>	<b>0.895</b>	0.848	0.889	0.877	0.845
2.500	1.0 -1.5	0.013	0.111	0.463	0.602	0.687	<b>0.810</b>	0.730	0.668	<b>0.756</b>	0.687	0.660
3.000	0.5 -1.0	0.006	0.059	0.203	0.292	0.318	<b>0.433</b>	0.350	0.284	<b>0.377</b>	0.318	0.275
4.000	0.5 -1.5	0.004	0.044	0.231	0.348	0.419	<b>0.591</b>	0.487	0.435	<b>0.558</b>	0.419	0.416
Mean power		0.083	0.293	0.447	0.500	0.458	<b>0.505</b>	<b>0.519</b>	0.469	0.481	0.482	0.457
p-value for (11,2,2;7,7,6)		0.228	0.073	0.038	0.025	0.028	0.028	0.037	0.069	0.080	0.031	0.080

Table 3. Pairwise comparisons of 11 tests for  $4 \times 7 = 28$  values of  $\theta$ , where each entry is the number of parameter values (out of 28 considered in the power calculations) for which the test to the left (defining the row) had greater power than the test above (defining the column).

	$\varphi_0$	$\varphi_{0,100}$	$\varphi_{0.5}$	$\varphi_{0.5,100}$	$\varphi_1$	$\varphi_{1,100}$	$\varphi_{0.5,1}$	$\varphi_A$	$\varphi_{COM(L)}$	$\varphi_S$	$\varphi_{CH}$	Total
$\varphi_0$	-	<b>0</b>	1	<b>0</b>	4	3	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	8
$\varphi_{0,100}$	<b>28</b>	-	7	6	10	9	7	7	9	9	9	101
$\varphi_{0.5}$	27	21	-	<b>0</b>	14	12	6	14	13	12	17	136
$\varphi_{0.5,100}$	<b>28</b>	<b>22</b>	<b>28</b>	-	18	15	13	21	18	17	21	201
$\varphi_1$	24	18	14	10	-	<b>0</b>	<b>0</b>	19	14	<b>0</b>	20	119
$\varphi_{1,100}$	25	19	16	13	<b>28</b>	-	17	20	21	19	20	198
$\varphi_{0.5,1}$	<b>28</b>	21	22	15	<b>28</b>	11	-	25	23	<b>28</b>	<b>28</b>	<b>229</b>
$\varphi_A$	<b>28</b>	21	14	7	9	8	3	-	10	8	<b>28</b>	136
$\varphi_{COM(L)}$	<b>28</b>	19	15	10	14	7	5	18	-	12	20	148
$\varphi_S$	<b>28</b>	19	16	11	<b>28</b>	9	<b>0</b>	20	16	-	21	168
$\varphi_{CH}$	<b>28</b>	19	11	7	8	8	<b>0</b>	<b>0</b>	8	7	-	96
Total	272	179	144	79	161	82	<b>51</b>	144	132	112	184	

row of Table 2, they are dominated at  $\alpha = 0.05$  by their corresponding adaptive tests  $\varphi_{0,0,100}$  ( $\alpha = 0.040$ ),  $\varphi_{0.5,100}$  ( $\alpha = 0.044$ ), and  $\varphi_{1,0,100}$  ( $\alpha = 0.039$ ). This is not surprising, and will be the case quite generally. Note that  $\varphi_{0.5,1}$  maximizes the average power, at 0.519, or the area under the power curve. The non-adaptive tests did not fare as well. Among the omnibus tests ( $\varphi_A$ ,  $\varphi_{COM(L)}$ ,  $\varphi_S$ , and  $\varphi_{CH}$ ),  $\varphi_{0.5,1}$  maximizes the power for 22 of the 28  $\theta$  values ( $\varphi_A$  and  $\varphi_{COM(L)}$  each maximize the power for three  $\theta$  values). Also,  $\varphi_{0.5,1}$  ( $p = 0.037$ ) and  $\varphi_S$  ( $p = 0.031$ ) are the only omnibus tests to yield statistical significance at  $\alpha = 0.05$  for  $\{(11,2,2);(7,7,6)\}$ . Table 3, above, shows that  $\varphi_{0.5,1}$  dominates both  $\varphi_S$  and  $\varphi_{CH}$ , and almost dominates  $\varphi_A$  and  $\varphi_{COM(L)}$  too, and does dominate them when  $\delta(\theta)$  is near the  $\delta_P$  value of 0.5 used by  $\varphi_{0.5,1}$ . In fact, only where  $\delta(\theta) \leq -0.5$  or  $\delta(\theta) \geq 2.5$  is  $\varphi_A$  or  $\varphi_{COM(L)}$  more powerful than  $\varphi_{0.5,1}$ . Among pairwise comparisons,  $\varphi_{0.5,1}$  has larger power than its competitor (each of the other ten tests are considered for each of 28 alternatives) for 229 out

of 280 comparisons, and 104 of the 112 comparisons to omnibus tests. The non-adaptive tests did not fare as well, but  $\varphi_S$  attained 168/280 or 57/112, respectively, which is quite respectable.

### Conclusion

In an effort to improve the comparison of two treatments on the basis of ordinal data, a new class of adaptive tests was defined, and shown to be admissible, while providing unambiguous p-values and a non-iterative construction. If one is interested in testing for  $\theta_1 > 0$ , and has no particular preference for any subset of  $\Omega_A$  relative to any other, then  $\varphi_{CH}$  would be a fine test to use.

However,  $\varphi_A$  and  $\varphi_{0.5,1}$  are also excellent omnibus tests, and are easier to compute than  $\varphi_{CH}$ . If one is interested in testing for stochastic order, and uses  $\theta_1 > 0$  only as a surrogate, then  $\varphi_A$  and  $\varphi_{0.5,1}$  are probably better tests than  $\varphi_{CH}$ . Certainly if one is in the situation treated in this article, with a preferred direction, then an appropriate adaptive test would be the test of choice. There is nothing particular about ordered trinomial distributions that makes this problem especially amenable to treatment with the adaptive

approach. For any hypothesis testing problem with a composite alternative hypothesis, one can enumerate the alternatives and the corresponding MP test for each. One can then apply each of these MP tests to a given outcome, and find the smallest of the resulting p-values. Using this minimized MP p-value as a test statistic produces a test analogous to  $\phi_A$ , and reduces to the uniformly most powerful test if one exists. If not, then the adaptive tests that bridge the gap between  $\phi_A$  and the MP tests to detect a favored direction should have good properties in a variety of contexts.

#### References

- Berger, V. W. (1998). Admissibility of exact conditional tests of stochastic order. *Journal of Statistical Planning and Inference*, 66, 39-50.
- Berger, V. W. (2000). Pros and cons of permutation tests. *Statistics in Medicine*, 19, 1319-1328.
- Berger, V. W., & Ivanova, A. (2002). The bias of linear rank tests when testing for stochastic order in ordered categorical data. *Journal of Statistical Planning and Inference*, 107, 237-247.
- Berger, V. W., Lunneborg, C., Ernst, M. D., Levine, J. G. (2002), Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods*, 1, 74-82.
- Berger, V. W., Permutt, T., & Ivanova A. (1998). The Convex hull test for ordered categorical data. *Biometrics*, 54, 1541-1550.
- Berger, V. W., & Sackrowitz, H. (1997). Improving tests for superior treatments in contingency tables. *Journal of American Statistical Association*, 92, 700-705.
- Chiara, S., Compora, E., Merlini, L., et al. (1993). Recurrent ovarian carcinoma: salvage treatment with platinum in patients responding to first-line platinum-based regimens. *European Journal of Cancer*, 29A, 652.
- Cohen, A., & Sackrowitz, H. (1998). Directional tests for one-sided alternatives in multivariate models. *Annals of Statistics*, 26, 2321-2338.
- Donegani, M. (1991). An adaptive and powerful randomization test. *Biometrika*, 78, 930-933.
- Edgington, E. S. (1995). *Randomization tests*. (3<sup>rd</sup> ed.). New York: Marcel Dekker.
- Frick, H. (2000). Undominated p-values and property *c* for unconditional one-sided two-sample binomial tests. *Biometrical Journal*, 42, 715-728.
- Gastwirth, J. L. (1985). The use of maximum efficiency robust tests in combining contingency tables and survival analysis. *The Journal of the American Statistical Association*, 80, 380-384.
- Good, P. (1994). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. New York: Springer-Verlag.
- Gross, S. T. (1981). On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *Journal of American Statistical Association*, 76, 935-941.
- Graubard, B. I., & Korn, E. L. (1987). Choice of column scores for testing independence in ordered 2xk contingency tables. *Biometrics*, 43, 471-476.
- Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of American Statistical Association*, 69, 909-923.
- Ivanova, A., & Berger, V. W. (2001). Drawbacks of integer scoring of ordered categorical data. *Biometrics*, 57, 567-570.
- Marden, J. I. (1991). Sensitive and sturdy p-values. *Annals of statistics*, 19, 918-934.
- McCullagh, P. (1980). Regression methods for ordinal data. *Journal of Royal Statistical Society, B* 42, 109-142.
- Permutt, T., & Berger, V. W. (2000). Rank tests in ordered 2xk contingency tables. *Communications in Statistics, Theory and Methods*, 29, 989-1003.
- Podgor, M. J., Gastwirth, J. L., Mehta, C. R. (1996). Efficiency robust tests of independence in contingency tables with ordered categories. *Statistics in Medicine*, 15, 2095-2105.
- Rohmel, J. & Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal*, 41, 149-170.
- Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Statistics*, 24, 220-238.

Rukhin, AL, Mak, KS (1992). Adaptive Test Statistics and Bahadur Efficiency. *Statistica Sinica*, 2, 541-552.

Appendix

Lemmas (with Proofs), and Proofs of Theorems

*Lemma 1.* Let  $c \in \Gamma$  and  $k \in \{0, 1, \dots, K_c\}$ . If  $|v_k(c) \pm \varepsilon(c)| < \infty$  then  $v_k(c) \pm \varepsilon(c) \notin V(c)$ . If  $v \in (v_k(c), v_{k+1}(c))$ , then  $M_v(c) = M_{v_{(k+1)}(c)}(c) - B_{v_{(k+1)}(c)}^-(c) = M_{v_{(k)}(c)}(c) - B_{v_{(k)}(c)}^+(c)$ .

*Proof.* Increasing (decreasing)  $v$  by  $\varepsilon(c)$  moves  $B_v^-(c)$  ( $B_v^+(c)$ ) into the interior of, and  $B_v^+(c)$  ( $B_v^-(c)$ ) completely out of, the new critical region, but if  $v \in V(c)$ , then no points of  $\Gamma - M_v(c)$  are moved into the new critical region (Table 1). Hence,  $o_{v-\varepsilon(c)}(c) = o_{v+\varepsilon(c)}(c) = 1$ , and neither  $v_k(c) - \varepsilon(c)$  nor  $v_k(c) + \varepsilon(c)$  is in  $V(c)$ . If  $v \notin V(c)$ , say  $v_k(c) < v < v_{k+1}(c)$ , then  $o_v(c) = 1$ , so  $B_v^+(c) = B_v^-(c) = \emptyset$  and  $M_v(c)$  will not change when  $v$  varies within  $(v_k(c), v_{k+1}(c))$ .

*Lemma 2.* If  $\delta_P \in \mathfrak{R}^1$ ,  $\tau > 0$ ,  $v_* \in v_{[\delta_P, \tau]}(c)$ , and  $v^* \in v^*(c)$ , then  $|\delta_P - v_*| \leq |\delta_P - v^*|$ .

*Proof.* If there exist  $v^* \in v^*(c)$  and  $v_* \in v_{[\delta_P, \tau]}(c)$  such that  $|\delta_P - v^*| < |\delta_P - v_*|$ , then  $p_{v^*}(c)(1 + |\delta_P - v^*|)^\tau < p_{\min(v_*)}(c)(1 + |\delta_P - v_*|)^\tau$ , and  $v_*$  cannot be in  $v_{[\delta_P, \tau]}(c)$ .

*Lemma 3.* For any  $\delta_P$ ,  $\tau > 0$ , and  $c \in \Gamma$ ,  $v_{[\delta_P, \tau]}(c) \subset V(c) \cup \delta_P$ .

*Proof.* Assume there exists  $v \neq \delta_P$  in  $v_{[\delta_P, \tau]}(c) - V(c)$ , say  $v_k(c) < v < v_{k+1}(c)$ . Let  $v^* = v_k(c)$  if  $\delta_P \leq v_k(c)$ ,  $v^* = \delta_P$  if  $v_k(c) < \delta_P < v_{k+1}(c)$ , or  $v^* = v_{k+1}(c)$  if  $v_{k+1}(c) \leq \delta_P$ . Now  $v^* \subset V(c) \cup \delta_P$  and

$$p_{\min(v)}(c)(1 + |\delta_P - v|)^\tau > p_{\min(v^*)}(c)(1 + |\delta_P - v^*|)^\tau.$$

*Lemma 4.* For any  $\delta_P$  and  $c \in \Gamma$ ,  $v_{[\delta_P, \tau]}(c) = \{\delta_P\}$  for sufficiently large  $\tau$ .

*Proof.* Let  $D_c(\delta_P) = \min_{v \in V(c) - \delta_P} |\delta_P - v| > 0$ . For  $\tau > 0$ , let  $v \in v_{[\delta_P, \tau]}(c) - \delta_P$ . By Lemma 3,  $v \in V(c) - \delta_P$ , so  $|\delta_P - v| \geq D_c(\delta_P)$ . If  $\tau > -\ln(p_{\min(\delta_P)}(c))/\ln(1 + D_c(\delta_P))$ , then  $p_{\min(v)}(c)(1 + |\delta_P - v|)^\tau \geq p_{\min(v)}(c)(1 + |D_c(\delta_P)|)^\tau > 1$ , contradicting  $v \in v_{[\delta_P, \tau]}(c)$ .

*Proof of Theorem 1.* By Theorem 3.3 of Berger (1998), it suffices to show that for any  $B \subset \Gamma$ , if  $c^*$  minimizes  $A(\delta_P, \tau, c)$  over  $B$ , then  $c^* \in D[B]$ . If  $c^* \notin D[B]$ , then  $c^*$  cannot, for any  $v$ , uniquely minimize  $p_v$  over  $B$ , and for every  $v$  there exists  $c \in B - c^*$  such that  $p_v(c) \leq p_v(c^*)$ . If  $v \notin V(c^*)$ , then  $o_v(c^*) = 1$ , so  $p_v(c) \neq p_v(c^*)$ , and  $p_v(c) \leq p_v(c^*) - \min_{c \in \Gamma} P_0\{c|\Gamma\}$ . Let  $v_1 \in v_{[\delta_P, \tau]}(c^*)$ . By the continuity in  $v$  of the function  $(1 + |\delta_P - v|)^\tau$ , one can, for any  $\varepsilon > 0$ , choose  $v_2 \notin V(c^*)$  suitably close to  $v_1$  to satisfy  $p_{v_2}(c^*) = p_{\min(v_1)}(c^*)$ , and, thus,

$$\begin{aligned} A(\delta_P, \tau, c) &= \min_{-\infty \leq v \leq \infty} [p_{\min(v)}(c)(1 + |(\delta_P - v)|)^\tau] \leq \\ & p_{v_2}(c)(1 + |(\delta_P - v_2)|)^\tau \\ & \leq [p_{v_2}(c^*) - \min_{c \in \Gamma} P_0\{c|\Gamma\}](1 + |\delta_P - v_2|)^\tau \\ & = [p_{\min(v_1)}(c^*) - \min_{c \in \Gamma} P_0\{c|\Gamma\}](1 + |\delta_P - v_2|)^\tau \\ & < A(\delta_P, \tau, c^*) - \min_{c \in \Gamma} P_0\{c|\Gamma\}(1 + |\delta_P - v_2|)^\tau + \varepsilon < A(\delta_P, \tau, c^*), \end{aligned}$$

the last inequality holding for  $\varepsilon < \min_{c \in \Gamma} P_0\{c|\Gamma\}$ . This is a contradiction.



## Within Groups Multiple Comparisons Based On Robust Measures Of Location

Rand R. Wilcox  
Dept of Psychology  
University of Southern California

H. J. Keselman  
Dept of Psychology  
University of Manitoba

---

Consider the problem of performing all pair-wise comparisons among  $J$  dependent groups based on measures of location associated with the marginal distributions. It is well known that the standard error of the sample mean can be large relative to other estimators when outliers are common. Two general strategies for addressing this problem are to trim a fixed proportion of observations or empirically check for outliers and remove (or down-weight) any that are found. However, simply applying conventional methods for means to the data that remain results in using the wrong standard error. Methods that address this problem have been proposed, but among the situations considered in published studies, no method has been found that gives good control over the probability of a Type I error when sample sizes are small (less than or equal to thirty); the actual probability of a Type I error can drop well below the nominal level. The paper suggests using a slight generalization of a percentile bootstrap method to address this problem.

Key words: M-estimators, trimming, bootstrap.

---

### Introduction

Outliers (unusually small or large values) can inflate the standard error of the sample mean which in turn can result in relatively poor power, and outliers can distort the sample mean resulting in a misleading representation of the typical response (e.g., Rosenberger & Gasko, 1983; Staudte & Sheather, 1990; Wilcox, 2001). When dealing with measures of location, two general strategies have been proposed for dealing with this problem.

The first is to simply trim a fixed proportion of the extreme values. In terms of maintaining a relatively low standard error under normality yet deal with situations where outliers are rather common, a 20% trimmed mean is often recommended (which is formally defined in the next section of this paper). The other strategy is to empirically check for outliers and remove (or downweight) any that are found. Various textbooks recommend some variation of the latter strategy and often refer to this as data cleaning.

If outliers are removed and the values are not erroneous (merely unusually large or small), applying standard methods for means to the remaining data results in using the wrong standard error, which in turn means poor control over the probability of a Type I error and inaccurate confidence intervals. Effective methods for dealing with this problem were derived for a range of situations, but when comparing measures of location associated with the marginal distributions of dependent groups, practical problems remain. Methods that avoid Type I error probabilities well above the nominal level are available, but when empirically checking and discarding outliers, the actual probability of a Type I error can drop well below the nominal level.

---

Rand R. Wilcox is a Professor of Psychology, a fellow of the Royal Statistical Society and the American Psychological Society, has published over 170 journal articles, and has recently written his fifth book on statistics. E-mail him at [rwilcox@usc.edu](mailto:rwilcox@usc.edu). H. J. Keselman is a Professor of Psychology, a fellow of the American Psychological Association and the American Psychological Society, and has published over 100 journal articles and book chapters related to the analysis of repeated measurements, multiple comparison procedures, and robust estimation and testing. E-mail him at [kesel@ms.umanitoba.ca](mailto:kesel@ms.umanitoba.ca).

For  $J$  dependent groups, let  $\theta_j$  be some measure of location associated with the  $j$ th marginal distribution. More formally, this paper is concerned with all pairwise comparisons where for every  $j < k$ , the goal is to test

$$H_0 : \theta_j = \theta_k. \quad (1)$$

Of particular interest is controlling the family-wise error rate (FWE), meaning the probability of at least one Type I error. When the sample size is small and the goal is to have FWE equal to .05, extant simulation results indicate that it is possible to ensure FWE will not exceed .05 by a substantial amount using 20% trimmed means in conjunction with a generalization of the bootstrap method (Wilcox, 1997b). A concern, however, is that the actual FWE can drop well below the nominal level suggesting that the method might have relatively low power.

Wilcox (1997b) also found that when using an estimator that in effect discards outliers (called a one-step M-estimator with Huber's  $\Psi$ ), poor control over FWE is obtained with sample sizes less than or equal to thirty. Currently, no method has been found that performs reasonably well in simulations when using this particular M-estimator and the sample size is small. So a practical issue remains: Is it possible to find a method that, in simulations, not only avoids FWE rates larger than the nominal level, it ensures that FWE will not be substantially below the nominal level when extreme values are discarded. This paper describes such a method which is based on a slight generalization of the percentile bootstrap.

#### Description of the Robust Estimators

The focus is on three measures of location. The first is a 20% trimmed mean. Generally, trimmed means simply remove a fixed proportion of the extreme observations. By fixed proportion is meant that the amount of trimming is not determined empirically by, for example, checking to see what proportion of the observations are outliers. The median and mean are trimmed means that represent the two extremes of the maximum amount and least amount of trimming, respectively. The choice of 20% trimming provides reasonably good efficiency under normality and it maintains relatively high

efficiency in situations where the sample mean performs poorly (Rosenberger & Gasko, 1983; Wilcox, 1997a), so we focus on it here. The 20% trimmed mean removes the smallest 20% of the observations, as well as the largest 20%, and averages the values that remain. If  $X_1, \dots, X_n$  is a random sample, let  $X_{(1)} \leq \dots \leq X_{(n)}$  be the observations written in ascending order and let  $g$  be equal to  $.2n$  rounded down to the nearest integer. Then a 20% trimmed mean is

$$\bar{X}_t = \frac{1}{n-2g} \sum_{i=g+1}^{n-g} X_{(i)}.$$

However, 20% trimmed means in particular, and trimmed means in general, suffer from at least two practical concerns. First, the amount of trimming is assumed to be fixed in advance. If the amount of trimming is set at 20%, efficiency is reasonably good versus the mean under normality, but when sampling from a sufficiently heavy-tailed distribution, efficiency can be poor versus using more trimming or switching to some robust M-estimator of location. A second general concern is that typically trimmed means assume symmetric trimming. That is, the same proportion of observations are trimmed from both tails of an empirical distribution. When sampling from an approximately symmetric distribution, symmetric trimming seems reasonable, but asymmetric trimming might be more appropriate as the degree of skewness increases. Well known theoretical results indicate how to estimate the standard error of a trimmed mean when asymmetric trimming is used (e.g., Huber, 1981), but now unsatisfactory probability coverage can result when sample sizes are small (e.g., Wilcox, 1997a). Also, if the amount of trimming is empirically determined, and the standard error is estimated by conditioning on this amount of trimming, even poorer control over probability coverage can result.

The second measure of location is a particular robust M-estimator. Generally, robust M-estimators are more flexible than trimmed means in the sense that they empirically determine whether a value is unusually large or small and then such values are down weighted in some manner. The particular M-estimator of interest here is the one-step M-estimator based on Huber's  $\Psi$ :

$$\frac{1.28(MADN)(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2}, \quad (2)$$

where  $M$  is the usual median,  $MAD$  is the median of the values  $X_1 - M, \dots, X_n - M$ ,  $MADN = MAD/.6745$ ,  $i_1$  is the number of observations  $X_i$  such that  $(X_i - M) < -K(MADN)$ ,  $i_2$  is the number of observations  $X_i$  such that  $(X_i - M) > K(MADN)$ , and  $K$  is some constant usually chosen to achieve good properties under normality. (See, for example, Staudte and Sheather, 1990.) This estimator empirically determines whether an observation is an outlier, trims it, averages the values that remain, but with asymmetric trimming an adjustment is made based on a measure of scale,  $MAD$ . The adjustment based on  $MAD$  is a consequence of how the population value of the one-step  $M$ -estimator is defined. It is the value  $\theta$  satisfying

$$E \left[ \Psi \left( \frac{X - \theta}{MADN} \right) \right] = 0, \quad (3)$$

where  $\Psi(x) = \max[-K; \min(K; x)]$ . Equation (3) can be solved with the Newton-Raphson method and a single iteration of this technique yields (with  $K = 1.28$ ) equation (2). The choice  $K = 1.28$  provides good efficiency under normality and its finite sample breakdown point is .5, the highest possible value. (The finite sample breakdown point of an estimator is the smallest proportion of observations, which when altered, can drive the value of an estimator to plus or minus infinity.) However, when performing all pair-wise comparisons among  $J$  dependent groups based on this one-step  $M$ -estimator, none of the techniques examined by Wilcox (1997b) performed well in simulations. Moreover, situations arise where even the most successful method can have Type I error probabilities well below the nominal level.

The third measure of location considered here is a so-called modified one-step  $M$ -estimator (MOM). The MOM estimator belongs to the class of skipped estimators originally proposed by Tukey and studied by Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972). The idea is simple: Check for outliers, discard any that are

found, and then average the values that remain. The class of skipped estimators studied by Andrews et al. is based on a boxplot outlier detection rule which has a finite sample breakdown point of only .25. Here an outlier detection rule based on  $M$  and  $MADN$  is used instead resulting in a location estimator having a finite sample breakdown point of .5 as well. (Huber, 1993, argues that at a minimum, an estimator should have a finite sample breakdown point of at least .1.)

An apparent disadvantage of skipped estimators is that expressions for their standard errors are very complicated when sampling from an asymmetric distribution. One of the main points in this paper is that a variation of the percentile bootstrap method not only circumvents this problem, it provides good probability coverage in simulations where no effective method based on a robust  $M$ -estimator has been found.

The modified one-step  $M$ -estimator begins by declaring  $X_i$  an outlier if

$$\frac{.6745 |X_i - M|}{MAD} > K,$$

where  $K$  is adjusted so that efficiency is good under normality. (Outlier detection rules based on the sample mean and variance are known to be unsatisfactory, e.g., Wilcox, 2001, pp. 34-35.) Then MOM is given by

$$\hat{\theta} = \sum_{i=i_1+1}^{n-i_2} \frac{X_{(i)}}{n - i_1 - i_2}, \quad (4)$$

where now  $i_1$  ( $i_2$ ) is the number of observations less (greater) than the median that are declared outliers. Here,  $K = 2.24$  is used which is approximately equal to the square root of the .975 quantile of a chi-square distribution with one degree of freedom. This particular outlier detection rule is a special case of a general method suggested by Rousseeuw and van Zomeren (1990.) It is noted that this choice for  $K$  yields good efficiency under normality.

In particular, using simulations with 10,000 replications, we found that with  $K = 2.24$ , the standard error of the sample mean divided by

the standard error of  $\hat{\theta}$  is approximately .9 for  $n = 20(5)100$ . For  $n = 10$  and  $15$ , this ratio is .88.

The Proposed Method for Pair-wise Comparisons

Here,  $\hat{\theta}_j$  represents the estimate of the measure of location associated with  $j$ th marginal distribution. Let  $X_{ij}$ ,  $i = 1, \dots, n, j = 1, \dots, J$  represent a random sample of size  $n$  from some  $J$ -variate distribution. So for fixed  $j$  and when using a trimmed mean,  $\hat{\theta}_j$  would be the 20% trimmed mean associated with  $X_{1j}, \dots, X_{nj}$ , ignoring the other data.

First consider a basic percentile bootstrap method for testing (1) which stems from Liu and Singh (1997) as well as Hall (1986) and is applied as follows. Obtain bootstrap samples by resampling with replacement  $n$  rows from the  $n$  by  $J$  matrix of  $X_{ij}$  values. Repeat this process  $B$  times and let  $\hat{\theta}_{bj}^*$  be the bootstrap estimate of  $\theta_j$  based on the  $b$ th bootstrap sample,  $b = 1, \dots, B; j = 1, \dots, J$ . (Here,  $\theta_j$  represents the population value of any of the three estimators under consideration.) Let

$$p_{jk}^* = P(\hat{\theta}_j^* > \hat{\theta}_k^*)$$

based on a random bootstrap sample. Here this probability is estimated with  $\hat{p}_{jk}^*$ , the proportion of bootstrap samples having  $\theta_{bj}^* > \theta_{bk}^*$ . Then if  $H_0$  is true,  $\hat{p}_{jk}^*$  has, asymptotically, a uniform distribution, so reject if  $\min(\hat{p}_{jk}^*, 1 - \hat{p}_{jk}^*) \leq \alpha/2$ .

To control FWE, some type of sequentially rejective method can be used. Here consideration was given to the approach derived by Rom (1990) as well as Hochberg (1988) which are outlined below. A positive feature of the methods just outlined is that for all three measures of location, simulation estimates of the FWE were less than or equal to the nominal level for all of the situations described in our simulations. This is true when using the Rom or the Hochberg method. However, a negative feature when testing at the .05 level was that when using MOM or Huber's M-estimator, the estimated FWE was typically less than .05 by an unacceptable amount. In fact, estimates dropped below .01, particularly when the correlations among the variables are high.

An examination of the simulation results

indicated why this problem arose. When  $\hat{\theta}_j = \hat{\theta}_k$ , it should be the case that  $\hat{p}_{jk}^* = .5$ . Near equality was found when the correlation between  $X_{ij}$  and  $X_{ik}$  is close to zero, but as the correlation increased, the difference between  $E(\hat{p}_{jk}^*)$  and .5 increased as well.

This observation suggests the following modification. Set

$$D_{ij} = X_{ij} - \hat{\theta}_j.$$

That is, shift the data so that the null hypothesis is true. Obtain a bootstrap sample of size  $n$  from the  $D_{ij}$  values and let  $\hat{\theta}_{cj}^*$  be the resulting estimate of  $\theta_j$ . Repeat this process  $B$  times and let  $\hat{p}_{cjk}^*$  be the proportion of times  $\hat{\theta}_{cj}^*$  is greater than  $\hat{\theta}_{ck}^*$ . Set

$$\hat{p}_{ajk}^* = \hat{p}_{jk}^* - \lambda(\hat{p}_{cjk}^* - .5),$$

where  $\lambda$  is a constant to be determined. Then for fixed  $j$  and  $k$ , reject  $H_0 : \theta_j = \theta_k$  if  $\hat{p}_{ajk}^*$  is sufficiently large or small.

For convenience, set

$$\hat{p}_{mjk}^* = \min(\hat{p}_{ajk}^*, 1 - \hat{p}_{ajk}^*)$$

and assume the goal is to have FWE equal to  $\alpha$ . One approach to controlling FWE is to proceed along the lines in Hochberg (1988). Writing the  $C = (J^2 - J) / 2\hat{p}_{mjk}^*$  values as  $p_{m1}, \dots, p_{mC}$ , put these  $C$  values in ascending order yielding  $\hat{p}_{m(1)} \leq \dots \leq \hat{p}_{m(C)}$ . For any  $i = C, C-1, \dots, 1$ , if  $\hat{p}_{m(i)} \leq \alpha / 2(C - i + 1)$ , reject the corresponding hypothesis as well as all hypotheses having smaller  $\hat{p}_{m(i)}$  values.

Rom's (1990) method is applied in the same manner as Hochberg's technique, only  $\alpha / 2(C - i + 1)$  is replaced by a value tabled by Rom. Situations were found where Rom's method was a bit less satisfactory in avoiding FWE above the nominal level, so it is not considered further. Yet another approach was derived by Benjamini

and Hochberg (2000), but it is known that this method does not control FWE, so it is not considered here.

There remains the problem of choosing  $\lambda$ . The strategy was to determine an appropriate value under normality with all correlations equal to zero and all marginal distributions having a common variance. The reason for considering all correlations equal to zero was that when using a trimmed mean, MOM, or an M-estimator with Huber's  $\Psi$ , this was found to maximize the probability of at least one Type I error among all the situations considered in the next section. For  $n = 11$  and  $20$ , it was found that  $\lambda = .1$  gave good results when using MOM or the M-estimator considered here when used in conjunction with Hochberg's method, and as  $n$  increases, the term  $\lambda(\hat{p}_{cjk}^* - .5)$  becomes negligible. Using  $\lambda = 0$  results in FWE typically being less than the nominal level, but often it was far below the nominal level. As for 20% trimmed means,  $\lambda = 0$  performed well (no correction is needed) when using Hochberg.

### Results

The small-sample properties of the methods just described were studied for  $J = 4$  with simulations where observations were generated from a multivariate normal distribution via the IMSL (1987) subroutine RNMVN. Nonnormal distributions were generated using the g-and-h distribution (Hoaglin, 1985). That is, first generate  $Z_{ij}$  from a multivariate normal distribution and set

$$X_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp(hZ_{ij}^2 / 2).$$

For  $g = 0$  this last expression is taken to be

$$X_{ij} = Z_{ij} \exp(hZ_{ij}^2 / 2).$$

The case  $g = h = 0$  corresponds to a normal distribution. Setting  $g = 0$  yields a symmetric distribution, and as  $g$  increases, skewness increases as well. Heavy-tailedness increases with  $h$ . The values for  $g$  and  $h$  were taken to be  $(g, h) = (0, 0), (0, .5), (.5, 0)$  and  $(.5, .5)$ . Table 1 contains

skewness ( $\kappa_1$ ) and kurtosis ( $\kappa_2$ ) values for the four g-and-h distributions used in the simulations.

Table 1: Some properties of the g-and-h distribution

g	h	$\kappa_1$	$\kappa_2$	$\hat{\kappa}_1$	$\hat{\kappa}_2$
0.0	0.0	0.00	3.00	0.00	3.0
0.0	0.5	0.00	—	0.00	11,896.2
0.5	0.0	1.75	8.9	1.81	9.7
0.5	0.5	—	—	120.10	18,393.6

When  $h > 1/k$ ,  $E(X - \mu)^k$  is not defined and the corresponding entry in Table 1 is left blank. A possible criticism of simulations performed on a computer is that observations are generated from a finite interval, so the moments are finite even when in theory they are not, in which case observations are not being generated from a distribution having the theoretical skewness and kurtosis values listed in Table 1. In fact, as  $h$  gets large, there is an increasing difference between the theoretical and actual values for skewness and kurtosis. Accordingly, Table 1 also lists the estimated skewness ( $\hat{\kappa}_1$ ) and kurtosis ( $\hat{\kappa}_2$ ) values based on 100,000 observations generated from the distribution. Simulations were also run where the marginal distributions were lognormal or exponential.

Simulations were run where the marginal distributions had equal and unequal variances. When working with skewed distributions, the marginal distributions were first shifted so that they have a  $\theta$  value of zero, and for the unequal variance case the  $i$ th observation in the  $j$ th group was multiplied by  $\sigma_j$ ,  $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 3, 4, 5)$ . That is, for skewed distributions, before multiplying the  $X_{ij}$  by  $\sigma_j$ , the observations were shifted by subtracting the population value of  $\theta$  so that when multiplying by  $\sigma_j$ , the null hypothesis remains true.

Five patterns of correlations were used. Four of the five correlation matrices have a common correlation,  $\rho$ , with  $\rho = 0, .1, .5$  and  $.8$ . The fifth correlation matrix had  $\rho_{12} = .8, \rho_{13} = .5, \rho_{14} = .2, \rho_{23} = .5, \rho_{24} = .2$  and  $\rho_{34} = .2$ . The largest and smallest estimates of FWE consistently occurred with the first and latter two correlation matrices, so for brevity, only the results for the first and fifth matrices are reported. These two correlation matrices are labeled C1 and C2, respectively.

Table 2 contains the estimated probability of at least one Type I error when using the multiple comparison procedure described in the previous section. The results are based on 2,000 replications. As is evident, reasonably good control over the probability of a Type I error is achieved. The main difficulty is that when using MOM, there are two instances where the estimate drops below .02.

Conclusion

The main point is that currently, no method for comparing robust measures of location associated with the marginal distributions is very satisfactory in simulations with small sample sizes. The results reported here illustrate that by using a slight generalization of the percentile bootstrap method, good control over the probability of a Type I error can be achieved in a wide range of situations when outliers are removed.

As for trimmed means, a basic (unmodified) percentile bootstrap method performs well. The three estimators used in Table 2 are designed to have reasonably good efficiency under normality, they have high efficiency when sampling from a heavy-tailed distribution where the sample mean performs poorly, so comparing groups as described would seem to have practical value. The M-estimator and modified M-estimator seem particularly attractive, and now it appears that a viable method for performing all pair-wise comparisons, based on the measures of location associated with the marginal distributions, is available when sample sizes are small.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances*. Princeton University Press, Princeton, NJ.

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25, 60-83.

Table 2: Estimated Type I error probabilities for g-and-h distributions,  $n = 11, J = 4$

g	h	Correlation	$\sigma$	Method		
				MOM	M-EST	$\bar{X}_t$
0.0	0.0	C1	(1,1,1,1)	.030	.020	.036
		C2		.067	.053	.051
0.0	0.5	C1	(1,1,1,1)	.040	.037	.044
		C2		.014	.015	.027
0.5	0.0	C1	(1,1,1,1)	.056	.059	.048
		C2		.025	.035	.033
0.5	0.5	C1	(1,1,1,1)	.041	.044	.042
		C2		.016	.023	.026
0.0	0.0	C1	(1,3,4,5)	.053	.051	.036
		C2		.057	.049	.034
0.0	0.5	C1	(1,3,4,5)	.041	.044	.038
		C2		.037	.048	.028
0.5	0.0	C1	(1,3,4,5)	.050	.058	.042
		C2		.056	.046	.034
0.5	0.5	C1	(1,3,4,5)	.041	.051	.041
		C2		.041	.048	.033

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20, 1803-1827.

Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics*, 14, 1431-1452.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g and h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring data tables, trends, and shapes*. (p. 461-515). New York: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800-802.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Huber, P. (1993). Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti & W. Stahel (Eds.) *New directions in statistical data analysis and robustness*. Boston: Birkhauser Verlag.

IMSL (1987). *Library I, vol. II*. Houston: International Mathematical and Statistical Libraries.

Liu, R. Y. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92, 266-277.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77, 663-666.

Rosenberger, J. L., & Gasko, M. (1983). In D. C. Hoaglin, F. Mosteller and J. W. Tukey (Eds.) *Understanding robust and exploratory data analysis*. New York: Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association*, 85, 633-639.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Wilcox, R. R. (1997a). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1997b). Pairwise comparisons using trimmed means or M-estimators when working with dependent groups. *Biometrical Journal*, 39, 677-688.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.

## Trimming, Transforming Statistics, And Bootstrapping: Circumventing the Biasing Effects Of Heteroscedasticity And Nonnormality

H. J. Keselman  
Dept. of Psychology  
University of Manitoba

Rand R. Wilcox  
Dept. of Psychology  
University of Southern  
California

Abdul R. Othman  
Universiti Sains  
Malaysia

Katherine Fradette  
University of Manitoba

---

Researchers can adopt different measures of central tendency and test statistics to examine the effect of a treatment variable across groups (e.g., means, trimmed means, M-estimators, & medians. Recently developed statistics are compared with respect to their ability to control Type I errors when data were nonnormal, heterogeneous, and the design was unbalanced: (1) a preliminary test for symmetry which determines whether data should be trimmed symmetrically or asymmetrically, (2) two different transformations to eliminate skewness, (3) the accuracy of assessing statistical significance with a bootstrap methodology was examined, and (4) statistics that use a robust measure of the typical score that empirically determined whether data should be trimmed, and, if so, in which direction, and by what amount were examined. The 56 procedures considered were remarkably robust to extreme forms of heterogeneity and nonnormality. However, we recommend a number of Welch-James heteroscedastic statistics which are preceded by the Babu, Padmanaban, and Puri (1999) test for symmetry that either symmetrically trimmed 10% of the data per group, or asymmetrically trimmed 20% of the data per group, after which either Johnson's (1978) or Hall's (1992) transformation was applied to the statistic and where significance was assessed through bootstrapping. Close competitors to the best methods were found that did not involve a transformation.

Key words: Symmetric vs. asymmetric trimming, Heteroscedastic statistic, Transformations to eliminate skewness, Preliminary test for symmetry, Bootstrapping.

---

### Introduction

#### Circumventing the Biasing Effects of Heteroscedasticity and Nonnormality

Developing new methods for locating treatment effects in the one-way independent groups design is a very active area of study. Much of the work centers on comparing measures of the

typical score when group variances are unequal and/or when data are obtained from nonnormal distributions. This continues to be an important area of work because the classical method of analysis, e.g., the analysis of variance F-test, is known to be adversely affected by heterogeneous group variances and/or nonnormal data. In particular, these conditions usually result in distorted rates of Type I error and/or a loss of statistical power to detect effects. Wilcox and Keselman (2002) discuss why this is so.

Many treatises have appeared on the topic of substituting robust measures of central tendency such as 20% trimmed means or M-estimators for the usual least squares estimator, i.e., the (least squares) means. Indeed, many investigators have demonstrated that one can achieve better control over Type I errors when robust estimators are substituted for least squares estimators in a heteroscedastic statistic such as Johanson's (1980) Welch-James (WJ)-type test (See e.g., Guo & Luh, 2000; Keselman, Kowalchuk, & Lix, 1998;

---

H. J. Keselman is Professor of Psychology, and fellow of the American Psychological Association and the American Psychological Society. He has published over 100 journal articles and book chapters. Email: [kesel@ms.umanitoba.ca](mailto:kesel@ms.umanitoba.ca). Rand R. Wilcox is Professor of Psychology. Email: [rwilcox@usc.edu](mailto:rwilcox@usc.edu). Katherine Fradette is an undergraduate honors student in the Department of Psychology. Abdul Rahman Othman is a lecturer in the School of Distance Education. Work on this project was supported by a grant by the National Sciences and Engineering Council of Canada.



Keselman, Lix, & Kowalchuk, 1998; Keselman, Wilcox, Taylor & Kowalchuk, 2000; Lix & Keselman, 1998; Luh & Guo, 1999; Wilcox, 1995, 1997; Wilcox, Keselman & Kowalchuk, 1998).

Another development in this area was to apply a transformation to a heteroscedastic statistic to eliminate the biasing effects of skewness. Indeed, Luh and Guo (1999) and Guo and Luh (2000) demonstrated that better Type I error control was possible when transformations (Hall's, 1992, or Johnson's, 1978, method) were applied to the WJ statistic with trimmed means.

Despite the advantages of using (20%) trimmed means, a heteroscedastic statistic with 20% trimming suffers from at least two practical concerns. First, situations arise where the proportion of outliers exceeds the percentage of trimming adopted, meaning that more trimming or some other measure of location, that is relatively unaffected by a large proportion of outliers, is needed. Second, if a distribution is highly skewed to the right, say, then at least in some situations it seems more reasonable to trim more observations from the right tail than from both tails.

Thus, using a heteroscedastic statistic with robust estimators, with or without transforming the statistic, may still not provide the best Type I error control. Two solutions that we consider in this paper are using a preliminary test for symmetry in order to determine whether data should be trimmed from both tails (symmetric trimming) or just from one tail (asymmetric trimming) and whether an estimator, other than the trimmed mean, that is, one that does not fix the amount of trimming a priori but empirically determines the amount and direction, or even the need for trimming, can provide better Type I error control.

The prevalent method of trimming is to remove outliers from each tail of the distribution of scores. In addition, the recommendation is to trim 20% from each tail (See Rosenberger & Gasko, 1983; Wilcox, 1995). However, asymmetric trimming has been theorized to be potentially advantageous when the distributions are known to be skewed, a situation likely to be realized with behavioral science data (See De Wet & van Wyk, 1979; Micceri, 1989; Tiku, 1980, 1982; Wilcox, 1994, 1995). Indeed, if a researcher's goal is to adopt a measure of the typical score, that is, a score that is representative of the bulk of the observations, then theory

certainly indicates that he/she should trim just from the tail in which outliers are located in order to get a score that represents the bulk of the observations; trimming symmetrically in this circumstance would eliminate representative scores, scores similar to the bulk of observations.

A stumbling block to adopting asymmetric versus symmetric trimming has been the inability of researchers to determine when to adopt one form of trimming over the other. That is, previous work has not identified a procedure which reliably identifies when data are positively or negatively skewed, rather than symmetric; thus researchers have not been able to successfully adopt one method of trimming versus the other. However, work by Hogg, Fisher and Randles (1975), later modified by Babu, Padmanaban, and Puri (1999), may provide a successful solution to this problem and accordingly enable researchers to successfully adopt asymmetric trimming in cases where it is needed thus providing them with measures of the typical score which more accurately corresponds to the bulk of the observations. The by-product of correctly identifying and eliminating only the outlying values should result in better Type I error control for heteroscedastic statistics that adopt trimmed means.

A concomitant issue that needs to be resolved is knowing how the 20% rule should be applied when trimming just from one tail. That is, should 40% of the longer tail of scores be trimmed since in total that amount is trimmed when trimming 20% in each tail? Or, should just 20% be trimmed from the one tail of the distribution? As well, the 20% rule is not universally recommended; others have had success with values other than 20%. For example, Babu et al. (1999) obtained good Type I error control, for the procedures they investigated, with 15% symmetric trimming. Indeed, as Huber (1993) argues, an estimator should have a breakdown point of at least .1; thus, even 10% trimming might provide effective Type I error control.

A second approach to the problem of direction and amount of trimming would be to adopt another robust estimator that does not a priori set the amount of trimming. Wilcox and Keselman (in press) introduced a modified M-estimator which empirically determines whether to trim symmetrically or asymmetrically and by what amount, or whether no trimming at all is

appropriate. In the context of a correlated groups design, they showed that their estimator does indeed provide effective Type I error control.

A last refinement that we will examine is the use of the bootstrap for hypothesis testing. Bootstrap methods have two practical advantages. First, theory and empirical findings indicate that they can result in better Type I error control than nonbootstrap methods (See Guo & Luh, 2000; Keselman, Kowalchuk, & Lix, 1998; Keselman, Lix, & Kowalchuk, 1998; Keselman, Wilcox, Taylor & Kowalchuk, 2000; Lix & Keselman, 1998; Luh & Guo, 1999; Wilcox (1995, 1997); Wilcox, Keselman & Kowalchuk, 1998). Second, certain variations of the bootstrap method do not require explicit expressions for standard errors of estimators. This makes hypothesis testing in some settings more flexible when other robust estimators (soon to be discussed) are used instead of trimmed means.

Thus, the purpose of our investigation was to compare rates of Type I error for numerous versions of the WJ heteroscedastic statistic versus two test statistics that use the estimator introduced by Wilcox and Keselman (2002). Variations of the WJ statistic will be based on asymmetric versus symmetric trimming, the amount of trimming, transformations of WJ and bootstrap versus nonbootstrap versions.

## Methods

### The WJ Statistic

Methods that give improved power and better control over the probability of a Type I error can be formulated using a general linear model perspective. Lix and Keselman (1995) showed how the various Welch (1938, 1951) statistics that appear in the literature for testing omnibus main and interaction effects as well as focused hypotheses using contrasts in univariate and multivariate independent and correlated groups designs can be formulated from this perspective, thus allowing researchers to apply one statistical procedure to any testable model effect. We adopt their approach in this paper and begin by presenting, in abbreviated form, its mathematical underpinnings.

A general approach for testing hypotheses of mean equality using an approximate degrees of freedom solution is developed using matrix

notation. The multivariate perspective is considered first; the univariate model is a special case of the multivariate. Consider the general linear model:

$$\mathbf{Y} = \mathbf{X}\beta + \xi, \quad (1)$$

where  $\mathbf{Y}$  is an  $N \times p$  matrix of scores on  $p$  dependent variables or  $p$  repeated measurements,  $N$  is the total sample size,  $\mathbf{X}$  is an  $N \times r$  design matrix consisting entirely of zeros and ones with  $\text{rank}(\mathbf{X}) = r$ ,  $\beta$  is an  $r \times p$  matrix of nonrandom parameters (i.e., population means), and  $\xi$  is an  $N \times p$  matrix of random error components. Let  $\mathbf{Y}_j$  ( $j = 1, \dots, r$ ) denote the submatrix of  $\mathbf{Y}$  containing the scores associated with the  $n$  subjects in the  $j^{\text{th}}$  group (cell) (For the one-way design considered in this paper  $n = n_j$ ). It is typically assumed that the rows of  $\mathbf{Y}$  are independently and normally distributed, with mean vector  $\beta_j$  and variance-covariance matrix  $\Sigma_j$  [i.e.,  $N(\beta_j, \Sigma_j)$ ], where the  $j^{\text{th}}$  row of  $\beta$ ,  $\beta_j = [\mu_{j1} \cdots \mu_{jp}]$ , and  $\Sigma_j \neq \Sigma_{j'}$  ( $j \neq j'$ ). Specific formulas for estimating  $\beta$  and  $\Sigma_j$ , as well as an elaboration of  $\mathbf{Y}$  are given in Lix and Keselman (1995, see their Appendix A).

The general linear hypothesis is

$$H_0 : \mathbf{R}\mu = \mathbf{0}, \quad (2)$$

where  $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^T$ ,  $\mathbf{C}$  is a  $df_C \times r$  matrix which controls contrasts on the independent groups effect(s), with  $\text{rank}(\mathbf{C}) = df_C \leq r$ , and  $\mathbf{U}$  is a  $p \times df_U$  matrix which controls contrasts on the within-subjects effect(s), with  $\text{rank}(\mathbf{U}) = df_U \leq p$ , ' $\otimes$ ' is the Kronecker or direct product function, and ' $^T$ ' is the transpose operator. For multivariate independent groups designs,  $\mathbf{U}$  is an identity matrix of dimension  $p$  (i.e.,  $\mathbf{I}_p$ ). The  $\mathbf{R}$  contrast matrix has  $df_C \times df_U$  rows and  $r \times p$  columns. In Equation 2,  $\mu = \text{vec}(\beta^T) = [\beta_1 \dots \beta_r]^T$ . In other words,  $\mu$  is the column vector with  $r \times p$  elements obtained by stacking the columns of  $\beta^T$ . The  $\mathbf{0}$  column vector is of order  $df_C \times df_U$ . (See Lix & Keselman, 1995, for illustrative examples.)

The generalized test statistic given by Johansen (1980) is

$$T_{WJ} = (\mathbf{R}\hat{\boldsymbol{\mu}})^T (\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1} (\mathbf{R}\hat{\boldsymbol{\mu}}), \quad (3)$$

where  $\hat{\boldsymbol{\mu}}$  estimates  $\boldsymbol{\mu}$ , and  $\hat{\boldsymbol{\Sigma}} = \text{diag}[\hat{\Sigma}_1/n_1 \dots \hat{\Sigma}_r/n_r]$ , a block matrix with diagonal elements  $\hat{\Sigma}_r/n_r$ . This statistic, divided by a constant,  $c$  (i.e.,  $T_{WJ}/c$ ), approximately follows an F distribution with degrees of freedom  $v_1 = df_C \times df_U$ , and  $v_2 = v_1(v_1 + 2)/(3A)$ , where  $c = v_1 + 2A - (6A)/(v_1 + 2)$ . The formula for the statistic,  $A$ , is provided in Lix and Keselman (1995).

When  $p = 1$ , that is, for a univariate model, the elements of  $\mathbf{Y}$  are assumed to be independently and normally distributed with mean  $\mu_j$  and variance  $\sigma_j^2$  [i.e.,  $N(\mu_j, \sigma_j^2)$ ]. To test the general linear hypothesis,  $\mathbf{C}$  has the same form and function as for the multivariate case, but  $\mathbf{U} = 1$ ,  $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1 \dots \hat{\mu}_r]^T$  and  $\hat{\boldsymbol{\Sigma}} = \text{diag}[\hat{\sigma}_1^2/n_1 \dots \hat{\sigma}_r^2/n_r]$ . (See Lix & Keselman's, 1995, Appendix A for further details of the univariate model.)

**Robust Estimation**

In this paper we apply robust estimates of central tendency and variability to the  $T_{WJ}$  statistic. That is, heteroscedastic ANOVA methods are readily extended to the problem of comparing trimmed means. The goal is to determine whether the effect of a treatment varies across  $J$  ( $j=1, \dots, J$ ) groups; that is, to determine whether a typical score varies across groups. When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the  $\mu_s$ . That is, to test the omnibus hypothesis in a one-way completely randomized design, the null hypothesis would be

$$H_0 : \mu_{t1} = \mu_{t2} = \dots = \mu_{tJ}.$$

Let  $Y_{(1)j} \leq Y_{(2)j} \leq \dots \leq Y_{(n_j)j}$  represent the ordered observations associated with the  $j^{\text{th}}$  group. Let  $g_j = [\gamma n_j]$ , where  $\gamma$  represents the proportion of observations that are to be trimmed in each tail of the distribution and  $[x]$  is the

greatest integer  $\leq x$ . The effective sample size for the  $j^{\text{th}}$  group becomes  $h_j = n_j - 2g_j$ . The  $j^{\text{th}}$  sample trimmed mean is

$$\mu_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)j}. \quad (4)$$

Wilcox (1995) suggested that 20% trimming should be used. (See Wilcox, 1995 and his references for a justification of the 20% rule.)

The sample Winsorized mean is necessary and is computed as

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}, \quad (5)$$

where

$$\begin{aligned} X_{ij} &= Y_{(g_j+1)j} \quad \text{if } Y_{ij} \leq Y_{(g_j+1)j} \\ &= Y_{ij} \quad \text{if } Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ &= Y_{(n_j-g_j)j} \quad \text{if } Y_{ij} \geq Y_{(n_j-g_j)j}. \end{aligned}$$

The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of a trimmed mean, is then given by

$$\hat{\sigma}_{wj}^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2. \quad (6)$$

The standard error of the trimmed mean is estimated with

$$\sqrt{(n_j - 1)\hat{\sigma}_{wj}^2/[h_j(h_j - 1)]}.$$

Under asymmetric trimming, and assuming, without loss of generality, that the distribution is positively skewed so that trimming takes place in the upper tail, the  $j^{\text{th}}$  sample trimmed mean is

$$\hat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=1}^{n_j-g_j} Y_{(i)j},$$

and the  $j^{\text{th}}$  sample Winsorized mean is

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij},$$

where

$$\begin{aligned} X_{ij} &= Y_{ij} \quad \text{if } Y_{ij} < Y_{(n_j-g_j)j} \\ &= Y_{(n_j-g_j)j} \quad \text{if } Y_{ij} \geq Y_{(n_j-g_j)j}. \end{aligned}$$

The sample Winsorized variance is again defined as (given the new definition of  $\hat{\mu}_{wj}$ )

$$\hat{\sigma}_{wj}^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2,$$

and the standard error of the mean again takes its usual form (given the new definition of  $\hat{\mu}_{wj}$ ).

Thus, with robust estimation, the trimmed group means ( $\hat{\mu}_{tj}$ ) replace the least squares group means ( $\hat{\mu}_j$ ), the Winsorized group variances estimators ( $\hat{\sigma}_{wj}^2$ ) replace the least squares variances ( $\hat{\sigma}_j^2$ ), and  $h_j$  replaces  $n_j$  and accordingly one computes the robust version of  $T_{WJ}$ ,  $T_{WJt}$ . (See Keselman, Wilcox, & Lix, 2001; for another justification of adopting robust estimates see Rocke, Downs & Rocke, 1982).

### Bootstrapping

Now we consider how extensions of the ANOVA method just outlined might be improved. In terms of probability coverage and controlling the probability of a Type I error, extant investigations indicate that the most successful method, when using a 20% trimmed mean (or some M-estimator), is some type of bootstrap method.

Following Westfall and Young (1993), and as enumerated by Wilcox (1997), let  $C_{ij} = Y_{ij} - \hat{\mu}_{tj}$ ; thus, the  $C_{ij}$  values are the empirical distribution of the  $j^{\text{th}}$  group, centered so that the sample trimmed mean is zero. That is, the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample. The strategy

behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value.

For each  $j$ , obtain a bootstrap sample by randomly sampling with replacement  $n_j$  observations from the  $C_{ij}$  values, yielding  $Y_1^*, \dots, Y_{n_j}^*$ . Let  $T_{WJt}^*$  be the value of Johansen's (1980) test based on the bootstrap sample. Now we randomly sample (with replacement  $n_j$ ),  $B$  bootstrap samples from the shifted/centered distributions each time calculating the statistic  $T_{WJt}^*$ . The  $B$  values of  $T_{WJt}^*$  are put in ascending order, that is,  $T_{WJt(1)}^* \leq \dots \leq T_{WJt(B)}^*$ , and an estimate of an appropriate critical value is  $T_{WJt(a)}^*$ , where  $a = (1 - \alpha)B$ , rounded to the nearest integer. One will reject the null hypothesis of location equality (i.e.,  $H_0 : \mu_{t1} = \mu_{t2} = \dots = \mu_{tJ}$ ) when  $T_{WJt} > T_{WJt(a)}^*$ , where  $T_{WJt}$  is the value of the heteroscedastic statistic based on the original nonbootstrapped data. Keselman et al. (2001) illustrate the use of this procedure for testing both omnibus and sub-effect (linear contrast) hypotheses in completely randomized and correlated groups designs.

### Transformations for the Welch-James Statistic

Guo and Luh (2000) and Luh and Guo (1999) found that Johnson's (1978) and Hall's (1992) transformations improved the performance of several heteroscedastic test statistics when they were used with trimmed means, including the WJ statistic, in the presence of heavy-tailed and skewed distributions.

In our study we, accordingly, compared both approaches for removing skewness when applied to the  $T_{WJt}$  statistic. Let  $Y_{ij} = (Y_{1j}, Y_{2j}, \dots, Y_{n_jj})$  be a random sample from the  $j^{\text{th}}$  distribution. Let  $\hat{\mu}_{tj}$ ,  $\hat{\mu}_{wj}$  and  $\hat{\sigma}_{wj}^2$  be, respectively, the trimmed mean, Winsorized mean and Winsorized variance of group  $j$ . Define the Winsorized third central moment of group  $j$  as

$$\hat{\mu}_{3j} = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^3.$$

Let

$$\tilde{\sigma}_{wj}^2 = \frac{(n_j - 1)}{(h_j - 1)} \hat{\sigma}_{wj}^2,$$

$$\tilde{\mu}_{wj} = \frac{n_j}{h_j} \hat{\mu}_{3j},$$

$$q_j = \frac{\tilde{\sigma}_{wj}^2}{h_j},$$

$$w_{tj} = \frac{1}{q_j},$$

$$U_t = \sum_{j=1}^J w_{tj},$$

and

$$\hat{\mu}_t = \frac{1}{U_t} \sum_{j=1}^J w_{tj} \hat{\mu}_{tj}.$$

Guo (2000) defined a trimmed mean statistic with Johnson's transformation as:

$$T_{\text{Johnson}_j} = (\hat{\mu}_{tj} - \hat{\mu}_t) + \frac{\tilde{\mu}_{wj}}{6\tilde{\sigma}_{wj}^2 h_j} + \frac{\tilde{\mu}_{wj}}{3\tilde{\sigma}_{wj}^4} (\hat{\mu}_{tj} - \hat{\mu}_t)^2 \tag{7}$$

From Guo and Luh (2000) we can deduce that a trimmed mean statistic with Hall's (1992) transformation would be:

$$T_{\text{Hall}_j} = (\hat{\mu}_{tj} - \hat{\mu}_t) + \frac{\tilde{\mu}_{wj}}{6\tilde{\sigma}_{wj}^2 h_j} + \frac{\tilde{\mu}_{wj}}{3\tilde{\sigma}_{wj}^4} (\hat{\mu}_{tj} - \hat{\mu}_t)^2 + \frac{\tilde{\mu}_{wj}^2}{27\tilde{\sigma}_{wj}^8} (\hat{\mu}_{tj} - \hat{\mu}_t)^3 \tag{8}$$

Keselman et al. (2001) indicated that sample trimmed means, sample Winsorized variances and trimmed sample sizes can be substituted for the

usual sample means, variances and sample sizes in the  $T_{wj}$  statistic. That is,

$$T_{WJ} = \sum_{j=1}^J w_{tj} (\hat{\mu}_{tj} - \hat{\mu}_t)^2,$$

which, when divided by  $c$ , is distributed as an F variable with df of  $J - 1$  and

$$v = (J^2 - 1) \left[ 3 \sum_{j=1}^J \frac{(1 - w_{tj} / U_t)^2}{h_j - 1} \right]^{-1},$$

where

$$c = (J - 1) \left( 1 + \frac{2(J - 2)}{J^2 - 1} \sum_{j=1}^J \frac{(1 - w_{tj} / U_t)^2}{h_j - 1} \right).$$

Now we can define

$$T_{WJ_{\text{Johnson}}} = \sum_{j=1}^J w_{tj} (T_{\text{Johnson}_j})^2 \tag{9}$$

and

$$T_{WJ_{\text{Hall}}} = \sum_{j=1}^J w_{tj} (T_{\text{Hall}_j})^2, \tag{10}$$

Then  $T_{WJ_{\text{Johnson}}}$  and  $T_{WJ_{\text{Hall}}}$ , when divided by  $c$ , are also distributed as F variates with no change in degrees of freedom.

### A Preliminary Test for Symmetry

A stumbling block to adopting asymmetric versus symmetric trimming has been the inability of researchers to determine when to adopt one form of trimming over the other. Work by Hogg et al. (1975) and Babu et al. (1999), however, may provide a successful solution to this problem. The details of this method are presented in Othman, Keselman, Wilcox, and Fradette (2003).

### The One-Step Modified M-Estimator (MOM)

For  $J$  independent groups (this estimator can also be applied to dependent groups) consider the

MOM estimator introduced by Wilcox and Keselman (in press). In particular, these authors suggested modifying the well-known one-step M-estimator

$$\frac{1.28(\text{MADN}_j)(i_2 - i_1) + \sum_{i=i_1+1}^{n_j-i_2} Y_{(i)j}}{n_j - i_1 - i_2}, \quad (11)$$

by removing  $1.28(\text{MADN}_j)(i_2 - i_1)$ , where  $\text{MADN}_j = \text{MAD}_j / .6745$ ,  $\text{MAD}_j =$  the median of the values  $|Y_{ij} - \hat{M}_j|, \dots, |Y_{n_j j} - \hat{M}_j|$ ,  $\hat{M}_j$  is the median of the  $j^{\text{th}}$  group,  $i_1 =$  the number of observations where  $Y_{ij} - \hat{M}_j < 2.24(\text{MADN}_j)$ , and  $i_2 =$  the number of observations where  $Y_{ij} - \hat{M}_j > 2.24(\text{MADN}_j)$ . Thus, the modified M-estimator suggested by Wilcox and Keselman is

$$\hat{\theta}_j = \sum_{i=i_1+1}^{n_j-i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2}. \quad (12)$$

The MOM estimate of location is just the average of the values left after all outliers (if any) are discarded. The constant 2.24 is motivated in part by the goal of having a reasonably small standard error when sampling from a normal distribution. Moreover, detecting outliers with Equation 12 is a special case of a more general outlier detection method derived by Rousseeuw and van Zomeren (1990).

MOM estimators, like trimmed means, can be applied to test statistics to investigate the equality of this measure ( $\theta$ ) of the typical score across treatment groups. The null hypothesis is

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_J,$$

where  $\theta_j$  is the population value of MOM associated with the  $j^{\text{th}}$  group. Two statistics can be used. The first was a statistic mentioned by Schrader and Hettmansperger (1980), examined by He, Simpson and Portnoy (1990) and discussed by Wilcox (1997, p. 164). The test is defined as

$$H = \frac{1}{N} \sum_{j=1}^J n_j (\hat{\theta}_j - \hat{\theta})^2 \quad (14)$$

where  $N = \sum_j n_j$  and  $\hat{\theta} = \sum_j \hat{\theta}_j / J$ . To assess statistical significance a (percentile) bootstrap method can be adopted. That is, to determine the critical value one centers or shifts the empirical distribution of each group; that is, each of the sample MOMs is subtracted from the scores in their respective groups (i.e.,  $C_{ij} = Y_{ij} - \text{MOM}_j$ ).

As was the case with trimmed means, the strategy is to shift the empirical distributions with the goal of estimating the null distribution of H which yields an estimate of an appropriate critical value. Now one randomly samples (with replacement), B bootstrap samples from the shifted/centered distributions each time calculating the statistic H, which when based on a bootstrap sample, is denoted as  $H^*$ . The B values of  $H^*$  are put in ascending order, that is,  $H_{(1)}^* \leq \dots \leq H_{(B)}^*$ , and an estimate of an appropriate critical value is  $H_{(a)}^*$ , where  $a = (1 - \alpha)B$ , rounded to the nearest integer. One will reject the null hypothesis of location equality when  $H > H_{(a)}^*$ .

The second method of analysis presented can be obtained in the following manner (See Liu & Singh, 1997). Let

$$\delta_{jj'} = \theta_j - \theta_{j'} \quad (j < j') \quad (15)$$

Thus, the  $\delta_{jj'}$ s are the all possible pairwise comparisons among the J treatment groups.

Now, if all groups have a common measure of location, (i.e.,  $\theta_1 = \theta_2 = \dots = \theta_J$ ), then  $H_0 : \delta_{12} = \delta_{13} = \dots = \delta_{J-1,J} = 0$ . A boot-strap method can be used to assess statistical significance, but for this procedure the data does not need to be centered. In contrast to the first method, the goal is not to estimate the null distribution of some appropriate test statistic. Rather, bootstrap samples are obtained for the  $Y_{ij}$  values and one rejects if the zero vector is sufficiently far from the center of the bootstrap estimates of the delta values. Thus, bootstrap samples are obtained from the  $Y_{ij}$  values rather

than the  $C_{ij}$ s. For each bootstrap replication ( $B = 599$  is again recommended) one computes the robust estimators (i.e., MOM) of location (i.e.,  $\hat{\theta}_{jb}^*$ ,  $j = 1, \dots, J$ ;  $b = 1, \dots, B$ ) and the corresponding estimates of  $\delta_{jj'b}^*$  ( $\hat{\delta}_{jj'b}^* = \hat{\theta}_{jb}^* - \hat{\theta}_{j'b}^*$ ). The strategy is to determine how deeply  $\mathbf{0} = (0 \ 0 \dots 0)$  is nested within the bootstrap values  $\hat{\delta}_{jj'b}^*$ , where  $\mathbf{0}$  is a vector having length  $K = J(J-1)/2$ . This assessment is made by adopting a modification of Mahalanobis' distance statistic.

For notational convenience, we can rewrite the  $K$  differences  $\hat{\delta}_{jj'}$  as  $\hat{\Delta}_1, \dots, \hat{\Delta}_K$  and their corresponding bootstrap values as  $\hat{\Delta}_{kb}^*$  ( $k = 1, \dots, K$ ;  $b = 1, \dots, B$ ). Thus, let

$$\bar{\Delta}_k^* = \frac{1}{B} \sum_{b=1}^B \hat{\Delta}_{kb}^*$$

and

$$Z_{kb} = \hat{\Delta}_{kb}^* - \bar{\Delta}_k^* + \hat{\Delta}_k.$$

(Note the  $Z_{kb}$ s are shifted bootstrap values having mean  $\hat{\Delta}_k$ .) Now define

$$S_{kk'} = \frac{1}{B-1} \sum (Z_{kb} - \bar{Z}_k)(Z_{k'b} - \bar{Z}_{k'}), \quad (16)$$

where

$$\bar{Z}_k = \frac{1}{B} \sum_{b=1}^B Z_{kb}.$$

(Note: The bootstrap population mean of  $\bar{\Delta}_k^*$  is known and is equal to  $\hat{\Delta}_k$ .)

With this procedure, one next computes

$$D_b = (\hat{\Delta}_b^* - \hat{\Delta})\mathbf{S}^{-1}(\hat{\Delta}_b^* - \hat{\Delta})', \quad (17)$$

where  $\hat{\Delta}_b^* = (\hat{\Delta}_{1b}^*, \dots, \hat{\Delta}_{Kb}^*)$  and  $\hat{\Delta} = (\hat{\Delta}_1, \dots, \hat{\Delta}_K)$ . Accordingly,  $D_b$  measures how closely  $\hat{\Delta}_b^*$  is

located to  $\hat{\Delta}$ . If the null vector ( $\mathbf{0}$ ) is relatively far from  $\hat{\Delta}$  one rejects  $H_0$ . Therefore, to assess statistical significance, put the  $D_b$  values in ascending order ( $D_{(1)} \leq \dots \leq D_{(B)}$ ) and let  $a = (1 - \alpha)B$  (rounded to the nearest integer). Reject  $H_0$  if

$$T \geq D_{(a)}, \quad (18)$$

where

$$T = (\mathbf{0} - \hat{\Delta})\mathbf{S}^{-1}(\mathbf{0} - \hat{\Delta})'. \quad (19)$$

It is important to note that  $\theta_1 = \theta_2 = \dots = \theta_J$  can be true iff:

$$H_0 : \theta_1 - \theta_2 = \dots = \theta_{J-1} - \theta_J = 0.$$

(Therefore, it suffices to test that a set of  $K$  pairwise differences equal zero.) However, to avoid the problem of arriving at different conclusions (i.e., sensitivity to detect effects) based on how groups are arranged (if all MOMs are unequal), we recommend that one test the hypothesis that all pairwise differences equal zero.

### Empirical Investigation

Fifty-six tests for treatment group equality were compared for their rates of Type I error under conditions of nonnormality and variance heterogeneity in an independent groups design with four treatments. The procedures we investigated were:

Trimmed Means with Symmetric Trimming (No preliminary test for symmetry):

1.-3. WJ10(15)(20)-WJ with 10% (15%) (20%) trimming

4.-6. WJB10(15)(20)-10% (15%) (20%) trimming and bootstrapping

7.-9. WJJ10(15)(20)-10% (15%) (20%) trimming and Johnson's transformation

10.-12. WJJB10(15)(20)-10% (15%) (20%) trimming with Johnson's transformation and bootstrapping

13.-15 WJH10(15)(20)-10% (15%) (20%) trimming and Hall's transformation

16.-18 WJHB10(15)(20)-10% (15%) (20%) trimming and Hall's transformation and bootstrapping

WJ with Q Statistics: Symmetric and Asymmetric Trimming:

19.-21. WJ1010(1515)(2020)-WJ. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

22.-24. WJB1010(1515)(2020)-WJ with bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

25.-27. WJJ1010(1515)(2020)-WJ with Johnson's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

28.-30. WJJB1010(1515)(2020)-WJ with Johnson's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

31.-33. WJH1010(1515)(2020)-WJ with Hall's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

34.-36. WJHB1010(1515)(2020)-WJ with Hall's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 10% (15%) (20%) one sided trimming.

37.-39. WJ1020(1530)(2040)-WJ. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

40.-42. WJB1020(1530)(2040)-WJ with bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

43.-45. WJJ1020(1530)(2040)-WJ with Johnson's transformation. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

46.-48. WJJB1020(1530)(2040)-WJ with Johnson's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

49.-51. WJH1020(1530)(2040)-WJ with Hall's transformation. If data is symmetric use 10%

(15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

52.-54. WJHB1020(1530)(2040)-WJ with Hall's transformation and bootstrapping. If data is symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

Modified M-Estimators:

55. MOMH

56. MOMT

We examined: (a) the effect of using a preliminary test to determine whether data are symmetric or not in order to determine whether symmetric or asymmetric trimming should be adopted (we present in Appendix A a SAS/IML program that can be used to obtain the Q-statistics), (b) the percentage of symmetric (10%, 15% or 20%) and asymmetric (10%, 15%, 20%, 30% or 40%) trimming used, (c) the utility of transforming the WJ statistic with either Johnson's (1978) or Hall's (1992) transformation, (d) the utility of bootstrapping the data, and (e) the use of two statistics with an estimator (MOM) that empirically determines whether data should be symmetrically or asymmetrically trimmed and by what amount, allowing also for the option of no trimming.

Additionally, four other variables were manipulated in the study: (a) sample size, (b) pairing of unequal variances and group sizes, and (c) population distribution.

We chose to investigate an unbalanced completely randomized design containing four groups because previous research efforts pertained to this design (e.g., Lix & Keselman, 1998; Wilcox, 1988). The two cases of total sample size and the group sizes were  $N = 70$  (10, 15, 20, 25) and  $N = 90$  (15, 20, 25, 30). We selected our values of  $n_j$  from those used by Lix and Keselman (1998) in their study comparing omnibus tests for treatment group equality; their choice of values was, in part, based on having group sizes that others have found to be generally sufficient to provide reasonably effective Type I error control (e.g., see Wilcox, 1994). The unequal variances were in a 1:1:1:36 ratio. Unequal variances and unequal group sizes were both positively and negatively paired. For positive (negative) pairings, the group having the fewest number of observations was associated with the population having the smallest (largest) variance, while the



group having the greatest number of observations was associated with the population having the largest (smallest) variance. These conditions were chosen since they typically produce conservative (liberal) results.

With respect to the effects of distributional shape on Type I error, we chose to investigate nonnormal distributions in which the data were obtained from a variety of skewed distributions. In addition to generating data from a  $\chi_3^2$  distribution, we also used the method described in Hoaglin (1985) to generate distributions with more extreme degrees of skewness and kurtosis. These particular types of nonnormal distributions were selected since educational and psychological research data typically have skewed distributions (Micceri, 1989; Wilcox, 1994). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions, which were identified by Micceri on the robustness of Student's t test, and they found that only distributions with the most extreme degree of skewness (e.g.,  $\gamma_1 = 1.64$ ) affected the Type I error control of the independent sample t statistic. Thus, since the statistics we investigated have operating characteristics similar to those reported for the t statistic, we felt that our approach to modeling skewed data would adequately reflect conditions in which those statistics might not perform optimally.

For the  $\chi_3^2$  distribution, skewness and kurtosis values are  $\gamma_1 = 1.63$  and  $\gamma_2 = 4.00$ , respectively. The other nonnormal distributions were generated from the g and h distribution (Hoaglin, 1985). Specifically, we chose to investigate two g and h distributions: (a)  $g = .5$  and  $h = 0$  and (b)  $g = .5$  and  $h = .5$ , where g and h are parameters that determine the third and fourth moments of a distribution. To give meaning to these values it should be noted that for the standard normal distribution  $g = h = 0$ . Thus, when  $g = 0$  a distribution is symmetric and the tails of a distribution will become heavier as h increases in value. Values of skewness and kurtosis corresponding to the investigated values of g and h are (a)  $\gamma_1 = 1.75$  and  $\gamma_2 = 8.9$ , respectively, and (b)  $\delta_1 = \delta_2 = \text{undefined}$ . These values of skewness and kurtosis for the g and h distributions

are theoretical values; Wilcox (1997, p. 73) reports computer generated values, based on 100,000 observations, for these values--namely  $\gamma_1 = 1.81$  and  $\gamma_2 = 9.7$  for  $g = .5$  and  $h = 0$  and  $\hat{\gamma}_1 = 120.10$  and  $\gamma_2 = 18,393.6$  for  $g = .5$  and  $h = .5$ . Thus, the conditions we chose to investigate could be described as extreme. That is, they are intended to indicate the operating characteristics of the procedures under substantial departures from homogeneity and normality, with the premise being that, if a procedure works under the most extreme of conditions, it is likely to work under most conditions likely to be encountered by researchers.

In terms of the data generation procedure, to obtain pseudo-random normal variates, we used the SAS generator RANNOR (SAS Institute, 1989). If  $Z_{ij}$  is a standard unit normal variate, then  $Y_{ij} = \mu_j + \sigma_j \times Z_{ij}$  is a normal variate with mean equal to  $\mu_j$  and variance equal to  $\sigma_j^2$ . To generate pseudo-random variates having a  $\chi^2$  distribution with three degrees of freedom, three standard normal variates were squared and summed.

To generate data from a g- and h-distribution, standard unit normal variables were converted to random variables via

$$Y_{ij} = \frac{\exp(gZ_{ij})^{-1}}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right),$$

according to the values of g and h selected for investigation. To obtain a distribution with standard deviation  $\sigma_j$ , each  $Y_{ij}$  was multiplied by a value of  $\sigma_j$ . It is important to note that this does not affect the value of the null hypothesis when  $g = 0$  (See Wilcox, 1994, p. 297). However, when  $g > 0$ , the population mean for a g- and h-distributed variable is

$$\mu_{gh} = \frac{1}{g(1-h)^{1/2}} (e^{g^2/2(1-h)} - 1)$$

(See Hoaglin, 1985, p. 503.) Thus, for those conditions where  $g > 0$ ,  $\mu_{ij}$  was first subtracted from  $Y_{ij}$  before multiplying by  $\sigma_j$ . When working with MOMs,  $\theta_j$  was first subtracted from each observation (The value of  $\theta_j$  was obtained from

generated data from the respective distributions based on one million observations.). Specifically, for procedures using trimmed means, we subtracted  $\mu_{tj}$  from the generated variates under every generated distribution. Correspondingly, for procedures based on MOMs, we subtracted out  $\theta_j$  for all distributions investigated.

Lastly, it should be noted that the standard deviation of a g- and h-distribution is not equal to one, and thus the values reflect only the amount that each random variable is multiplied by and not the actual values of the standard deviations (See Wilcox, 1994, p. 298). As Wilcox noted, the values for the variances (standard deviations) more aptly reflect the ratio of the variances (standard deviations) between the groups. Five thousand replications of each condition were performed using a .05 statistical significance level. According to Wilcox (1997) and Hall (1986), B was set at 599; that is, their results suggest that it may be advantageous to chose B such that  $1 - \alpha$  is a multiple of  $(B + 1)^{-1}$ .

### Results

For previous investigations, when we have evaluated Type I error rates, we adopted Bradley's (1978) liberal criterion of robustness. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error ( $\hat{\alpha}$ ) must be contained in the interval  $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$ . Therefore, for the five percent level of statistical significance used in this study, a test would be considered robust in a particular condition if its empirical rate of Type I error fell within the interval  $.025 \leq \hat{\alpha} \leq .075$ .

Correspondingly, a test was considered to be nonrobust if, for a particular condition, its Type I error rate was not contained in this interval. We have adopted this standard because we felt that it provided a reasonable standard by which to judge robustness. That is, it has been our opinion that applied researchers should be comfortable working with a procedure that controls the rate of Type I error within these bounds, if the procedure limits the rate across a wide range of assumption violation conditions.

Type I error rates can be obtained from the first author's web site at the following address: [www.umanitoba.ca/faculties/arts/psychology](http://www.umanitoba.ca/faculties/arts/psychology). Based on this criterion of robustness, the procedures we investigated were remarkably robust to the cases of heterogeneity and nonnormality. That is, out of the 672 empirical values tabled (Tables 1-10) only 24, or approximately 3.5 percent of the values, did not fall within the .025-.075 interval (Values not falling in this interval are in boldface in the tables.)

Even though, in general, the procedures exhibited good Type I error control from the Bradley (1978) liberal criterion perspective, in the interest of making discriminations between the procedures, we went on to a second examination of the data adopting Bradley's stringent criterion of robustness. For this criterion, a statistic is considered robust, under a .05 significance level, if the empirical value falls in the interval .045-.055 (Non-bolded values not falling in this interval are underlined in the tables.). The tables as well contain information regarding the average Type I error rate and the number of empirical values not falling in the stringent interval for each procedure investigated; these values (excluding MOMH and MOMT values), along with the range of values over the 12 investigated conditions, are reproduced in summary form in Table 1.

Table 1. WJ Summary Statistics

<u>20% Symmetric Trimming</u>						
	<u>WJ20</u>	<u>WJJ20</u>	<u>WJH20</u>	<u>WJB20</u>	<u>WJJB20</u>	<u>WJHB20</u>
Range	.041-.079	.043-.075	.043-.076	.030-.047	.033-.047	.033-.047
Average	.058	.056	.056	.040	.041	.041
# of Nonrobust Values	12	9	9	10	9	10
<u>20% Symmetric and 40% Asymmetric Trimming</u>						
	<u>WJ2040</u>	<u>WJJ2040</u>	<u>WJH2040</u>	<u>WJB2040</u>	<u>WJJB2040</u>	<u>WJHB2040</u>
Range	.059-.084	.051-.077	.051-.079	.040-.053	.037-.053	.037-.052
Average	.071	.066	.068	.045	.048	.047
# of Nonrobust Values	12	11	11	4	2	2
<u>20% Symmetric and 20% Asymmetric Trimming</u>						
	<u>WJ2020</u>	<u>WJJ2020</u>	<u>WJH2020</u>	<u>WJB2020</u>	<u>WJJB2020</u>	<u>WJHB2020</u>
Range	.048-.075	.054-.071	.054-.072	.030-.051	.033-.055	.034-.054
Average	.059	.060	.060	.043	.047	.046
# of Nonrobust Values	8	9	9	6	4	4
<u>15% Symmetric Trimming</u>						
	<u>WJ15</u>	<u>WJJ15</u>	<u>WJH15</u>	<u>WJB15</u>	<u>WJJB15</u>	<u>WJHB15</u>
Range	.036-.067	.047-.067	.048-.067	.025-.047	.033-.048	.032-.048
Average	.051	.053	.054	.039	.042	.041
# of Nonrobust Values	8	4	4	9	8	8

Table 1. WJ Summary Statistics (continued)

15% Symmetric and 30% Asymmetric Trimming

	<u>WJ1530</u>	<u>WJJ1530</u>	<u>WJH1530</u>	<u>WJB1530</u>	<u>WJJB1530</u>	<u>WJHB1530</u>
Range	.057-.078	.050-.079	.050-.082	.035-.049	.041-.054	.039-.054
Average	.064	.063	.064	.045	.049	.048
# of Nonrobust Values	12	7	9	3	3	2

15% Symmetric and 15% Asymmetric Trimming

	<u>WJ1515</u>	<u>WJJ1515</u>	<u>WJH1515</u>	<u>WJB1515</u>	<u>WJJB1515</u>	<u>WJHB1515</u>
Range	.043-.065	.053-.072	.053-.073	.025-.045	.037-.050	.036-.050
Average	.053	.059	.060	.039	.046	.045
# of Nonrobust Values	7	8	8	9	4	5

10% Symmetric Trimming

	<u>WJ10</u>	<u>WJJ10</u>	<u>WJH10</u>	<u>WJB10</u>	<u>WJJB10</u>	<u>WJHB10</u>
Range	.038-.075	.053-.072	.055-.073	.025-.048	.033-.053	.033-.053
Average	.053	.059	.060	.039	.045	.043
# of Nonrobust Values	10	9	9	9	4	4

10% Symmetric and 20% Asymmetric Trimming

	<u>WJ1020</u>	<u>WJJ1020</u>	<u>WJH1020</u>	<u>WJB1020</u>	<u>WJJB1020</u>	<u>WJHB1020</u>
Range	.047-.075	.055-.072	.056-.074	.032-.052	.039-.057	.041-.057
Average	.059	.062	.063	.044	.049	.049
# of Nonrobust Values	8	11	12	5	2	2

Table 1. WJ Summary Statistics (continued)

10% Symmetric and 10% Asymmetric Trimming

	<u>WJ1010</u>	<u>WJJ1010</u>	<u>WJH1010</u>	<u>WJB1010</u>	<u>WJJB1010</u>	<u>WJHB1010</u>
Range	.038-.075	.055-.075	.056-.076	.023-.050	.033-.058	.032-.058
Average	.054	.064	.065	.039	.048	.042
# of Nonrobust Values	10	11	12	7	6	5

*Note:* Nonrobust values are those outside the interval .045-.055.

## Tests Based on MOMs

Of the 12 conditions examined, MOMH values ranged from .027 to .073, with an average value of .049; nine values fell outside of Bradley's (1978) stringent interval. MOMT values ranged from .014 to .060, with an average value of .038; six values fell outside the interval and most occurred when data were obtained from the  $g = .5$  and  $h = .5$  distribution. We describe our results predominately from Table 1; however, we, occasionally, also rely on the detailed information contained in the ten tables not contained in the paper.

## 20% Symmetric and 20% (40%) Asymmetric Trimming

Empirical results for 20% symmetric trimming conform to those reported in the literature. That is, the WJ test is generally robust with the liberal criterion of robustness, occasionally, however, resulting in a liberal rate of error (see Wilcox et al., 1998). Adopting a transformation for skewness improves rates of Type I error and further improvement is obtained when adopting bootstrap methods (see Luh & Guo, 1999). However, most of the values reported in the tables did not fall within the bounds of the stringent criterion. In particular, the number of these deviant values ranged from a low of 9 (WJJ20, WJH20, WJJB20) to a high of 12 (WJ20).

Keeping the total amount of trimmed values at 40%, regardless of whether data were trimmed symmetrically or asymmetrically, based on the preliminary test for symmetry, resulted in liberal rates of error, except when bootstrapping methods

were adopted. Indeed, when bootstrapping was adopted for assessing statistical significance and a transformation was/was not applied to the statistic (WJJB2040, WJHB2040, WJB2040), rates of Type I error were well controlled; the number of values falling outside the stringent interval were two, two and four, respectively, with corresponding average rates of error of .048, .047 and .045.

## 15% Symmetric and 15% (30%) Asymmetric Trimming.

Similar results were found to those previously reported, however, a few differences are noteworthy. First, none of the values fell outside the liberal criterion, though with the exception of WJJ15 and WJH15, the number of values outside of the stringent criterion was large, obtaining values of 8 and 9. Also noteworthy is that for 15% symmetric trimming bootstrapping did not result in improved rates of Type I error.

On the other hand, bootstrapping was quite effective for controlling errors when trimming was based on the preliminary test for symmetry and either 15% or 30% of the data were trimmed symmetrically or asymmetrically. Without bootstrapping, rates, on occasion, reached values above .075 and the number of values falling outside the stringent criterion ranged from 7 to 12. With bootstrapping, no value exceeded .075, in fact no value exceeded .054, and the number of values outside the stringent criterion was small--3 (WJB1530), 3 (WJJB1530) and 2 (WJHB1530).

When trimming was 15%-symmetric or 15%-asymmetric, based on the preliminary test for symmetry, again, all empirical values were contained in the liberal interval, ranging from a

low value of .025 (WJB1515) to a high value of .073 (WJH1515). However, the number of values falling outside the stringent interval varied over the tests examined, ranging from a low of 4 values (WJJB1515) to a high value of 9 values (WJB1515). The best two procedures were WJJB1515 (4 values outside the stringent criterion) and WJHB1515 (5 values outside the stringent criterion).

#### 10% Symmetric and 10% (20%) Asymmetric Trimming

Results are not generally dissimilar from those reported for the other two trimming rules. That is, when adopting a 10% symmetric rule, all rates were contained in the liberal interval, though with the 10% rule, bootstrapping and transforming the statistic for skewness was effective in limiting the number of deviant values (WJJB10 and WJHB10), while the remaining methods were not nearly as successful.

For 10% symmetric trimming or 20% asymmetric trimming, based on the preliminary test for symmetry, empirical rates were again best controlled when bootstrapping methods were applied. In particular, the number of deviant values ranged from 2 to 5, with fewer deviant values occurring when a transformation for skewness was applied to WJ (i.e., WJJB1020 and WJHB1020). The nonbootstrapped tests, on the other hand, frequently had rates falling outside the stringent interval; 8 for WJ1020 and 11 for WJJ1020 and WJH1020.

Adopting 10% symmetric or asymmetric trimming resulted in rates that generally also fell within the liberal criterion of Bradley (1978), except for two exceptions: .076 for WJH1010 and .023 for WJB1010. Once again, using a transformation to eliminate skewness and adopting bootstrapping to assess statistical significance resulted in relatively good Type I error control. That is, WJJB1010 and WJHB1010 had, respectively, 6 and 5 values falling outside the stringent interval, with corresponding average rates of error of .048 and .042.

#### Symmetric Trimming (10% vs 15% vs 20%).

Our last examination of the data was a comparison of the rates of Type I error across the various percentages of symmetric trimming. Only two liberal values (.076 and .079), according to the

.025-.075 criterion, were found across the three cases of symmetric trimming and they occurred under 20% symmetric trimming. The total number of values outside the .045-.055 criterion for 20%, 15% and 10% symmetric trimming were 58, 41 and 45, respectively; the corresponding average Type I error rates (across the six averages reported in the table) were .049, .047 and .050. The four procedures with the fewest values (i.e., 4) outside the stringent interval were WJJ15, WJH15, WJJB10 and WJHB10.

#### Discussion

In our investigation we examined various test statistics that can be used to compare treatment effects across groups in a one-way independent groups design. Issues that we examined were whether: (1) a preliminary test for symmetry can be used effectively to determine whether data should be trimmed symmetrically or asymmetrically when used in combination with a heteroscedastic statistic that compares trimmed means, (2) the amount of trimming affects error rates of these heteroscedastic statistics, (3) transformations to these heteroscedastic statistics improve results, (4) bootstrapping methodology provides yet additional improvements and (5) an estimator (MOM) that empirically determines whether one should trim, and, if so, by what amount and from which tail(s) of the distribution, can effectively control rates of Type I error, and how those rates compare to the other methods investigated.

We found that the fifty-six procedures examined performed remarkably well. Of the 672 empirical values, only 24, or approximately 3.5 percent of the values, did not fall within the bounds of .025-.075, a criterion that many investigators have used to assess robustness. Based on this criterion, only six procedures did not perform well--namely MOMT, WJ2040, WJJ2040, WJH2040, WJJ1530 and WJH1530; that is, they all had two or more values less than .025 or greater than .075. The vast majority of these nonrobust values occurred under our most extreme case of nonnormality:  $g = .5$  and  $h = .5$ .

On the basis of the more stringent criterion defined by Bradley (1978), five methods demonstrated exceptionally tight Type I error control. They were WJJB2040, WJHB2040, WJHB1530, WJJB1020 and WJHB1020. The

number of values not falling in the stringent interval was two for each procedure. In addition, the average rate of error was .048, .047, .048, .049 and .049, respectively. Common to these six procedures is the use of a transformation to eliminate skewness (either Hall's, 1992, or Johnson's, 1978) and the use of bootstrapping methodology to assess statistical significance. Two close competitors were the WJB1530 and WJJB1530 tests, each had three values outside .045-.055, with average rates of error of .045 and .049, respectively.

Based on our results we recommend WJJB1020 or WJHB1020; that is, the WJ heteroscedastic statistic which trims, based on a preliminary test for symmetry, 10% in each tail or 20% in one of the two tails and then transforms the test with a transformation to eliminate the effects of skewness (either Hall, 1978, or Johnson, 1992) and where statistical significance is determined from bootstrapping methodology. We recommend one of these methods, over the other three tests which also limited the number of discrepant values to two, because the other methods can result in greater numbers of data being discarded. It is our impression that applied researchers would prefer a method that compared treatment performance across groups with a measure of the typical score which was based on as much of the original data as possible--a very reasonable view. It is also worth mentioning that relatively good results are also possible by adopting a simpler WJ method--namely the WJ test with just bootstrapping. In particular, WJB1530 and WJB2040 resulted in 3 and 4 values outside the stringent interval and each had an average Type I error rate of .045.

Another noteworthy finding was that other percentages of symmetric trimming work better in the one-way design than 20% symmetric trimming. In particular, we found four methods involving less trimming than 20% (WJJ15, WJH15, WJJB10 and WJHB10) that provided good Type I error control, resulting in fewer values outside .045-.055 than identical procedures based on 20% trimming. For two of the methods (WJJ15 and WJH15), bootstrapping methodology is not required.

We conclude by reminding the reader that we examined fifty-six test statistics under conditions of extreme heterogeneity and nonnormality. Thus, we believe we have identified procedures that are

truly robust to cases of heterogeneity and nonnormality likely to be encountered by applied researchers and therefore we are very comfortable with our recommendation. That is, we believe we have found a very important result--namely, very good Type I error control is possible with relatively modest amounts of trimming.

We demonstrate the computations involved for obtaining the test of symmetry in Appendix A. We include this illustration, even though we provide software in Appendix A to obtain numerical results, because we believe it is instructive to see how Q2 and Q1 are obtained.

### References

- Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal*, 41(3), 321-339.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- De Wet, T., & Van Wyk, J. W. J. (1979). Efficiency and robustness of Hogg's adaptive trimmed means. *Communications in Statistics, Theory and Methods*, A8(2), 117-128.
- Guo, J. H., & Luh, W. M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, 49, 1-7.
- Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, 14, 1431-1452.
- Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B*, 54, 221-228.
- He, X., Simpson, D. G., & Portnoy, S. L. (1990). Breakdown robustness of tests. *Journal of the American Statistical Association*, 85, 446-452.
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g- and h-distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables, trends, and shapes* (p. 461-513). New York: Wiley.
- Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. *Journal of the American Statistical Association*, 70, 656-661.

- Huber, P. J. (1993). Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti, & W. Stahel (Eds.) *New directions in statistical data analysis and robustness*. Boston: Verlag.
- Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, *67*, 85-92.
- Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, *73*, 536-544.
- Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika*, *63*, 145-163.
- Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, *3*, 123-141.
- Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2001). A robust approach to hypothesis testing. Paper presented at the annual meeting of the Western Psychological Association, Maui, HI.
- Keselman, H. J., Wilcox, R. R., Taylor, J., & Kowalchuk, R. K. (2000). Tests for mean equality that do not require homogeneity of variances: Do they really work? *Communications in Statistics, Simulation and Computation*, *29*, 875-895.
- Liu, R. Y., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, *92*, 266-277.
- Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, *58*, 409-429 (58, 853).
- Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin*, *117*, 547-560.
- Luh, W., & Guo, J. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*, *52*, 303-320.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.
- Othman, A. R., Keselman, H. J., Wilcox, R. R., Fradette, K., & Padmanabhan, A. R. (2002). A test of symmetry. *Journal of Modern Applied Statistical Methods*, *1*(2), 310-315.
- Rocke, D. M., Downs, G. W., & Rocke, A. J. (1982). Are robust estimators really necessary? *Technometrics*, *24*(2), 95-101.
- Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers leverage points. *Journal of the American Statistical Association*, *85*, 633-639.
- Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians and trimean. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.). *Understanding robust and exploratory data analysis* (p. 297-336). New York: Wiley.
- SAS Institute Inc. (1989). *SAS/IML software: Usage and reference, version 6* (1st ed.). Cary, NC: Author.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error probabilities of the t test to departures from population normality. *Psychological Bulletin*, *111*, 352-360.
- Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance. *Biometrika*, *67*, 93-101.
- Tiku, M.L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. *Journal of Statistical Planning and Inference*, *4*, 123-143.
- Tiku, M.L. (1982). Robust statistics for testing equality of means and variances. *Communications in Statistics, Theory and Methods*, *11*(22), 2543-2558.
- Welch B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330-336.
- Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York: Wiley.



Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology*, *41*, 109-117.

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*, 289-306.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, *65*(1), 51-77.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.

Wilcox, R. R., & Keselman, H. J. (in press). Repeated measures one-way ANOVA based on a modified one-step M-estimator. *British Journal of Mathematical and Statistical Psychology*.

Wilcox, R. R., & Keselman, H. J. (2002). Some modern data analysis methods: Basics and recent developments. Manuscript submitted for publication.

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology*, *51*, 123-134.

#### Appendix A SAS/IML Program for Q-Statistics

```
*Checking for symmetry using the Q2 and Q1 indices presented in Babu,
Padmanabhan and Puri (1999);
*This program details all the steps in obtaining the Q2 and Q1 indices;
OPTIONS NOCENTER;
PROC IML;
RESET NONAME;
*Although the Q2 and Q1 calculations differ, both share common steps;
*Hence, they are incorporated into one module QMOD with the variable
QCHOICE being the switch that activates Q2 or Q1: 1 activates Q1 and 2
activates Q2;
START QMOD(QCHOICE,Y,OSY,GINFO,Q) GLOBAL(NY,WOBS,BOBS,PER);
  G = INT(PER#NY);
  NYPRIME = NY - 2#G;
  NPRIME = SUM(NYPRIME);
  *Initialize group information matrix;
  IF QCHOICE = 1 THEN GINFO = J(BOBS,8,0);
  ELSE IF QCHOICE = 2 THEN GINFO = J(BOBS,9,0);
  *Initialize for first pass;
  F = 1;
  M = 0;
  DO J = 1 TO BOBS;
    SAMP = NY[J];
    SAMPPR = NYPRIME[J];
    L = M + SAMP;
    YT = Y[F:L];
    TEMP = YT;
    *Sorting group elements in ascending order;
    YT[RANK(TEMP),] = TEMP;
    FIRST = G[,J] + 1;
    LAST = SAMP - G[,J];
    FPRIME = F + FIRST - 1;
    LPRIME = F + LAST - 1;
```

```

*Get group information;
GINFO[J,1] = J;      *Group number;
IF QCHOICE = 1 THEN DO;
  GINFO[J,2] = SAMPPR; *Possibly trimmed group size;
  GINFO[J,3] = FPRIME; *Starting position in possibly trimmed data
                    stream for group j;
  GINFO[J,4] = LPRIME; *Ending position in possibly trimmed data
                    stream for group j;
END; *if QCHOICE = 1;
ELSE IF QCHOICE = 2 THEN DO;
  GINFO[J,2] = SAMP; *Group size;
  GINFO[J,3] = F;   *Starting position in data stream for group j;
  GINFO[J,4] = L;   *Ending position in data stream for group j;
END; *if QCHOICE = 2;
*Calculating the mean of the upper and lower 5% of data in group j;
*This is common in both Q1 and Q2;
NJP05 = (LAST-FIRST+1)#0.05;
IF NJP05 <= 1 THEN DO;
  UP05J = YT[LAST];
  LP05J = YT[FIRST];
END; *if NJP05 <=1;
ELSE DO;
  A = INT(NJP05);
  FR = NJP05 - A;
  UP05 = YT[LAST-A+1:LAST];
  UP05J = (FR#YT[LAST-A] + SUM(UP05))/NJP05;
  LP05 = YT[FIRST:FIRST+A-1];
  LP05J = (SUM(LP05) + FR#YT[FIRST+A])/NJP05;
END; **if NJP05 > 1;
GINFO[J,5] = UP05J; *Upper 5% mean of group j;
GINFO[J,6] = LP05J; *Lower 5% mean of group j;
IF QCHOICE = 1 THEN DO;
  *Calculating the mean of the middle 50% of data in group j;
  *This calculation is done in Q1 only;
  NJP25 = (LAST-FIRST+1)#0.25;
  A = INT(NJP25);
  FR = NJP25 - A;
  ME = YT[FIRST+A+1:LAST-A-1];
  MIDJ = ((1-FR)#YT[FIRST+A] + SUM(ME) + (1-FR)#YT[LAST-A])/(2#NJP25);
  Q1J = (UP05J - MIDJ)/(MIDJ - LP05J);
  GINFO[J,7] = MIDJ; *Middle 50% mean of possibly trimmed group j;
  GINFO[J,8] = Q1J; *Q1 index of group j;
END; *if QCHOICE = 1;
IF QCHOICE = 2 THEN DO;
  *Calculating the mean of the upper and lower 50% of data in group j;
  *This calculation is done in Q2 only;
  NJP5 = (LAST-FIRST+1)#0.5;
  A = INT(NJP5);
  FR = NJP5 - A;
  UP5 = YT[LAST-A+1:LAST];
  UP5J = (FR#YT[LAST-A] + SUM(UP5))/NJP5;

```

```

LP5 = YT[FIRST:FIRST+A-1];
LP5J = (SUM(LP5) + FR#YT[FIRST+A])/NJP5;
Q2J = (UP05J - LP05J)/(UP5J - LP5J);
GINFO[J,7] = UP5J; *Upper 50% mean of group j;
GINFO[J,8] = LP5J; *Lower 50% mean of group j;
GINFO[J,9] = Q2J; *Q2 index of group j;
END; *if QCHOICE = 2;
*Update for next pass;
M = L;
F = F + NY[J];
IF J = 1 THEN OSY = YT;
ELSE OSY = OSY//YT;
END; *DO J;
IF QCHOICE = 1 THEN Q = SUM(GINFO[1:3,8]`#NYPRIME)/NPRIME;
ELSE IF QCHOICE = 2 THEN Q = SUM(GINFO[1:3,9]`#NYPRIME)/NPRIME;
FINISH; *QMOD;
START SHOWGRP(X, GINFO);
X1 = X[GINFO[1,3]:GINFO[1,4]]`;
X2 = X[GINFO[2,3]:GINFO[2,4]]`;
X3 = X[GINFO[3,3]:GINFO[3,4]]`;
PRINT 'GRP1:' X1[FORMAT=3.0];
PRINT 'GRP2:' X2[FORMAT=3.0];
PRINT 'GRP3:' X3[FORMAT=3.0];
FINISH; *SHOWGRP;
START Q2Q1AD;
PRINT 'DETAILED OUTPUT FOR THE Q-STATISTICS';
*Calculating Q2;
PER = 0; *Q2 does not require trimming of data;
QCHOICE = 2;
CALL QMOD(QCHOICE,Y,OSY,Q2INFO,Q2);
PRINT ;;
PRINT 'Y IN THE VARIOUS GROUPS';
CALL SHOWGRP(Y,Q2INFO);
PRINT ;;
PRINT 'ORDER STATISTICS OF Y';
CALL SHOWGRP(OSY,Q2INFO);
OUTQ2 = Q2INFO[,1:2]||Q2INFO[,5:9];
C1 = {"GRP" "GRP SIZE" "UP5% MEAN" "LO5% MEAN" "UP50% MEAN" "LO50% MEAN" "Q2J"};
PRINT ;;
PRINT 'INTERMEDIATE OUTPUTS FOR Q2';
PRINT OUTQ2[COLNAME=C1 FORMAT=10.4];
PRINT 'Q2 =' Q2[FORMAT=10.4];
IF Q2 < 3 THEN DO;
  PER = 0;
PRINT 'DATA DISTRIBUTION IS NORMAL-TAILED. USE ALL DATA TO DETERMINE Q1.';
END; *if Q2 < 3;
ELSE IF Q2 > 5 THEN DO;
  PER = 0.2;
PRINT 'DATA DISTRIBUTION IS VERY HEAVY-TAILED. DO 20% SYMMETRIC TRIMMING TO
DETERMINE Q1.';
END; *if Q2 > 5;

```

```

ELSE DO; *if 3 <= Q2 <= 5;
  PER = 0.1;
PRINT 'DATA DISTRIBUTION IS HEAVY-TAILED. DO 10% SYMMETRIC TRIMMING TO
DETERMINE Q1.';
END; *if 3 <= Q2 <=5;
*Calculating Q1;
QCHOICE = 1;
CALL QMOD(QCHOICE,Y,OSY,Q1INFO,Q1);
PRINT /;
PRINT 'ORDER STATISTICS OF POSSIBLY TRIMMED Y';
CALL SHOWGRP(OSY,Q1INFO);
OUTQ1 = Q1INFO[,1:2]||Q1INFO[,5:8];
C2 = {"GRP" "GRP SIZE" "UP5% MEAN" "LO5% MEAN" "MID50% MEAN" "Q1J"};
PRINT ;
PRINT 'INTERMEDIATE OUTPUTS FOR Q1';
PRINT OUTQ1[COLNAME=C2 FORMAT=10.4];
PRINT 'Q1 =' Q1[FORMAT=10.4];
IF Q1 < 0.5 THEN PRINT 'DATA DISTRIBUTION IS LEFT-SKEWED.';
ELSE IF Q1 > 2 THEN PRINT 'DATA DISTRIBUTION IS RIGHT-SKEWED.';
ELSE PRINT 'DATA DISTRIBUTION IS SYMMETRIC.'; *if 0.5 <= Q1 <= 2;
FINISH; *Q2Q1AD;
***INPUT DATA VECTOR;
*Data is purposely typed in the following manner to show where Groups 1-3
entries are;
*SAS treats this as a 35x1 column vector;
Y = {42, 40, 32, 48, 32, 52, 41, 35, 30, 99, 40, 35, 34, 39,
50, 49, 35, 43, 36, 40, 56, 41, 40, 64, 42,
48, 51, 63, 51, 60, 51, 83, 55, 55, 48};
*Group sizes are entries in the following 1x3 row vector;
NY = {15 10 10};
*WOBS and BOBS are variable names carried over from past programs;
*WOBS = within subjects groups;
WOBS = NCOL(Y);
*BOBS = between subject groups;
BOBS = NCOL(NY);
RUN Q2Q1AD;

```

-----

DETAILED OUTPUT FOR THE Q-STATISTICS  
Y IN THE VARIOUS GROUPS

GRP1: 42 40 32 48 32 52 41 35 30 99 40 35 34 39 50

GRP2: 49 35 43 36 40 56 41 40 64 42

GRP3: 48 51 63 51 60 51 83 55 55 48

ORDER STATISTICS OF Y

GRP1: 30 32 32 34 35 35 39 40 40 41 42 48 50 52 99

GRP2: 35 36 40 40 41 42 43 49 56 64

GRP3: 48 48 51 51 51 55 55 60 63 83

INTERMEDIATE OUTPUTS FOR Q2

GRP	GRP SIZE	UP5% MEAN	LO5%MEAN	UP50%MEAN	LO50% MEAN	Q2J
1	15	99	30	52.2667	34.2667	3.8333
2	10	64	35	50.8	38.4	2.3387
3	10	83	48	63.2	49.8	2.6119

$Q2 = 3.0573$

DATA DISTRIBUTION IS HEAVY-TAILED. DO 10% SYMMETRIC TRIMMING TO DETERMINE  $Q1$ .

ORDER STATISTICS OF POSSIBLY TRIMMED Y

GRP1: 32 32 34 35 35 39 40 40 41 42 48 50 52

GRP2: 36 40 40 41 42 43 49 56

GRP3: 48 51 51 51 55 55 60 63

INTERMEDIATE OUTPUTS FOR  $Q1$

GRP	GRP SIZE	UP5% MEAN	LO5% MEAN	MID50% MEAN	$Q1J$
1	13	52	32	38.8846	1.9050
2	8	56	36	41.5	2.6364
3	8	63	48	53	2

$Q1 = 2.1330$

DATA DISTRIBUTION IS RIGHT-SKEWED.

## A Test Of Symmetry

Abdul R. Othman  
Universiti Sains Malaysia

H. J. Keselman  
Dept. of Psychology  
University of Manitoba

Rand R. Wilcox  
Dept. of Psychology  
Univ. of Southern California

Katherine Fradette  
University of Manitoba

A. R. Padmanabhan  
Monash University, Australia

---

When data are nonnormal in form classical procedures for assessing treatment group equality are prone to distortions in rates of Type I error and power to detect effects. Replacing the usual means with trimmed means reduces rates of Type I error and increases sensitivity to detect effects. If data are skewed, say to the right, then it has been postulated that asymmetric trimming, to the right, should be better at controlling rates of Type I error and power to detect effects than symmetric trimming from both tails of the data distribution. Keselman, Wilcox, Othman and Fradette (2002) found that Babu, Padmanabhan and Puri's (1999) test for symmetry when combined with a heteroscedastic statistic which compared either symmetrically or asymmetrically determined means provided excellent Type I error control even when data were extremely heterogeneous and very nonnormal in form. In this paper, we present a detailed discussion of the Babu et al. procedure as well as a numerical example demonstrating its use.

Key words: Symmetry, Preliminary test

---

### Introduction

Keselman, Wilcox, Othman and Fradette (2002) found that by utilizing a test for symmetry prior to testing for equality of trimmed means they were able to achieve excellent Type I error control even though data were extremely heterogeneous and very nonnormal in form. In particular, they used a test for symmetry first proposed by Hogg, Fisher,

and Randles (1975) and subsequently modified by Babu, Padmanaban and Puri (1999) in order to determine whether data should be trimmed symmetrically or asymmetrically. Asymmetric trimming has been theorized to be potentially advantageous when the distributions are known to be skewed, a situation likely to be realized with behavioral science data (See De Wet & van Wyk, 1979; Micceri, 1989; Tiku, 1980, 1982; Wilcox, 1995). That is, theoretical considerations suggest that when data are say skewed to the right then in order to achieve robustness to nonnormality and greater sensitivity to detect effects one should trim data just from the upper tail of the data distribution. Indeed, Keselman et al. found that by combining a test for mean equality with a preliminary test for symmetry Type I error rates could be substantially improved for the nonnormal and heterogeneous distributions they examined. Because space considerations prevented them from providing a full description of the symmetry test we present the method herein and illustrate its application with a numerical example.

---

Abdul Rahman Othman is a lecturer in the School of Distance Education. His areas of interests are psychometrics and applied statistics. H. J. Keselman is Professor of Psychology. Email: [kesel@ms.umanitoba.ca](mailto:kesel@ms.umanitoba.ca). Rand R. Wilcox is Professor of Psychology. Email: [rwilcox@usc.edu](mailto:rwilcox@usc.edu). Katherine Fradette is an undergraduate honors student in the Department of Psychology. Her plans include graduate work in quantitative methods. A. R. Padmanabhan teaches in the Department of Mathematics in Australia. Work on this project was supported by the Natural Sciences and Engineering Council of Canada.

### Theoretical Background

The Babu et al. (1999) procedure is based, in part, on the work of Hogg et al. (1975). Specifically, for these authors, the hypothesis of interest was  $H_0: \theta = 0$  against  $H_A: \theta > 0$ , where  $\theta$  is the location parameter of interest. They proposed a test to detect the nature of the underlying distribution before proceeding with (nonparametric) tests of  $H_0$ .

In particular, they defined  $Y_1, Y_2, \dots, Y_m$  as a random sample from  $F(y)$ , and  $Y_{m+1}, Y_{m+2}, \dots, Y_n$  as a random sample from  $F(y - \theta)$ . Then  $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$  are the ordered statistics of the combined random samples and  $Y_{\text{med}}$  is the median of the combined samples.

Hogg et al.'s (1975) procedure to detect the nature of the underlying distribution is composed of two tests, a test of the heaviness of the tail of the distribution using the  $Q_2$  statistic and a test of symmetry using the  $Q_1$  statistic. Their work was based on papers by Uthoff (1970, 1973). Hogg et al. (1975) chose a test statistic enumerated by Uthoff (1973, Equation 2) as a basis to define their  $Q_2$  index. This index determined whether the tail of the underlying distribution is light or heavy. They first approximated it as

$$\frac{Y_{(n)} - Y_{(1)}}{2\sum |Y_{(i)} - Y_{\text{med}}| / n}.$$

They transformed this ratio into

$$Q_2 = \frac{(U_{0.05} - L_{0.05})}{(U_{0.5} - L_{0.5})},$$

where  $U_{0.05}$  and  $L_{0.05}$  are, respectively, the means of the upper and lower 5% of the order statistics of the sample and  $U_{0.5}$  and  $L_{0.5}$  are, respectively, the means of the upper and lower 50% of the order statistics of the combined sample.

Again, based on the work of Uthoff (1970, Equation 1), Hogg et al. (1975) derived their  $Q_1$  index:

$$Q_1 = \frac{(U_{0.05} - MID)}{(MID - L_{0.05})},$$

where MID is the mean of the middle 50% of the combined sample. Thus, this index determines the symmetry of the underlying distribution.

Babu et al. (1999) extended the use of these two indices to more than two groups. They proposed that both indices be calculated within the groups and weighted means of these indices be the overall estimates of  $Q_2$  and  $Q_1$ . They also proposed adjustments to the  $Q_1$  index whereby the amount of data needed to calculate the index depended on the outcome of the calculation of the  $Q_2$  index.

### Determination of Symmetry

Consider the problem of comparing distributions  $F_1 = F_2 = \dots = F_J$ . One way of approaching this problem is to consider the one-way ANOVA problem of comparing means  $\mu_1 = \mu_2 = \dots = \mu_J$  from  $J$  distributions  $F_1(y) = F(y - \mu_1)$ ,  $F_2(y) = F(y - \mu_2)$ ,  $\dots$ ,  $F_J(y) = F(y - \mu_J)$ . When the distributions are unknown and one cannot assume that they are normal with equal variances, Babu et al. (1999) suggested the following procedure to determine heavy-tailedness and symmetry prior to applying the appropriate test on the location parameters:

Let  $Y_{ij} = (Y_{1j}, Y_{2j}, \dots, Y_{n_jj})$  be a sample from an unknown distribution  $F_j$ . Let  $Y_{(1)j} \leq Y_{(2)j} \leq \dots \leq Y_{(n_j)j}$  represent the ordered observations associated with the  $j^{\text{th}}$  group. Let  $\gamma$  be the proportion of the data in the sample that are of interest as either the proportion of data to be trimmed or the proportion of data to be used in the calculation of several intermediate variables leading to two statistics, namely  $Q_2$  and  $Q_1$ . Let  $g = [\gamma n_j] + 1$ , where  $[x]$  represents the greatest integer less than  $\gamma n_j$  and  $r = g - \gamma n_j$ . It is important to note that trimming here, and the amount trimmed, is just for purposes of assessing symmetry.

### $Q_2$ Index

Prior to determining the symmetry of the distributions, the nature of their tails is examined. The  $Q_2$  index determines whether  $F_1(y)$ ,  $F_2(y)$ ,  $\dots$ ,  $F_J(y)$  are normal-tailed, heavy-tailed or very heavy-tailed. Tail classification is determined in the following manner:

1. Define  $U_{\gamma j}$  and  $L_{\gamma j}$  as the means of the upper and lower  $\gamma n_j$  order statistics, respectively, of the sample  $Y_j$ .

Case 1. If  $\gamma n_j \leq 1$ ,  
then  $U_{\gamma,j} = Y_{(n_j),j}$  and  $L_{\gamma,j} = Y_{(1),j}$ .

Case 2. If  $\gamma n_j > 1$   
then

$$U_{\gamma,j} = \frac{1}{\gamma n_j} \left( \sum_{i=n_j-\gamma+2}^{n_j} Y_{(i),j} + (1-r)Y_{(n_j-\gamma+1),j} \right) \text{ and}$$

$$L_{\gamma,j} = \frac{1}{\gamma n_j} \left( \sum_{i=1}^{\gamma-1} Y_{(i),j} + (1-r)Y_{(g),j} \right)$$

2. Calculate  $U_{0.05,j}$  and  $L_{0.05,j}$  as the mean of the upper and lower  $0.05n_j$  order statistics of  $Y_j$ , respectively.
3. Calculate  $U_{0.5,j}$  and  $L_{0.5,j}$  as the mean of the upper and lower  $0.5n_j$  order statistics of  $Y_j$ , respectively.
4. For each  $j$ , set  $Q_{2,j} = (U_{0.05,j} - L_{0.05,j}) / (U_{0.5,j} - L_{0.5,j})$ .
5. Using  $Q_{2,j}, j = 1, 2, \dots, J$ , from # 4 compute

$$Q_2 = \left( \sum_{j=1}^J n_j Q_{2,j} \right) / \left( \sum_{j=1}^J n_j \right)$$

6. If  $Q_2 < 3$  then  $F$  is classified as normal-tailed. If  $3 \leq Q_2 < 5$  then  $F$  is classified as heavy-tailed. If  $Q_2 \geq 5$  then  $F$  is classified as very heavy-tailed.

**Q<sub>1</sub> Index**

Once the nature of the tails of the distributions is known, the  $Q_1$  index, which determines the symmetry of the distributions, is calculated. To calculate the  $Q_1$  index one should:

1. Based on  $Q_2$ , determine the number of sample points in each sample  $Y_j$  to be used. Define this as  $n_j^*$ . (This is the Babu et al., 1999, modification of the Hogg et al., 1975, proposal for computing  $Q_1$ .) Specifically, if  $Q_2 < 3$  then use all sample points in  $Y_j$ . If  $3 \leq Q_2 < 5$  then trim the top and bottom 10% of the sample points and use the middle 80% in  $Y_j$ . If  $Q_2 \geq 5$  then trim the top and bottom 20% of the sample points and use the middle 60% in  $Y_j$ .
2. Let  $MID_j$  to be the mean of the middle 50% of the order statistics of the sample points in sample  $Y_j$  defined in #1. According to A. R. Padmanaban

(personal communication, June 26, 2001),  $MID_j$  is calculated in the following manner:

Discard the top and bottom 25% of the order statistics of  $Y_j$ .

The remainder is the middle 50% of the order statistics of  $Y_j$ .

Hence,  $g^* = [0.25n_j^*] + 1$  and  $r^* = g^* - 0.25n_j^*$ . Therefore,  $MID_j$  is given by

$$MID_j = \frac{1}{0.5n_j^*} \left[ \sum_{i=g^*}^{n_j^*-g^*} Y_{(i),j} + r^*(Y_{(g^*),j} + Y_{(n_j^*-g^*+1),j}) \right]$$

3. For each  $j$ , set

$$Q_{1,j} = (U_{0.05,j} - MID_j) / (MID_j - L_{0.05,j}).$$

4. Using  $Q_{1,j}, j = 1, 2, \dots, J$ , from # 3 compute

$$Q_1 = \left( \sum_{j=1}^J n_j^* Q_{1,j} \right) / \left( \sum_{j=1}^J n_j^* \right)$$

5. If  $Q_1 < 1/2$ ,  $F$  is deemed to be left skewed. If  $1/2 \leq Q_1 \leq 2$ , then  $F$  is considered to be symmetric. If  $Q_1 > 2$ , then  $F$  is designated as right skewed.

**Computational Example**

Suppose we want to test the null hypothesis,  $H_0: F_1(x) = F_2(x) = F_3(x)$  based on the following data set.

Table 1. Data set.

Groups	Order Statistics	$n_j$
1	30 32 32 34 35 35 39 40 40 41 42 48 50 52 99	15
2	35 36 40 40 41 42 43 49 56 64	10
3	48 48 51 51 51 55 55 60 63 83	10

Note: The tabled values were chosen so that the data would be classified as heavy-tailed.

**Calculating  $Q_2$  (Tail thickness)**

Notice that  $0.05n_j < 1$  for  $j = 1, 2, 3$ . Therefore,  $U_{0.05,1} = Y_{(15,1)} = 99$ ,  $U_{0.05,2} = Y_{(10,2)} = 64$ ,  $U_{0.05,3} = Y_{(10,3)} = 83$ , and  $L_{0.05,1} = Y_{(1,1)} = 30$ ,



$L_{0.05,2} = Y_{(1,2)} = 35$ , and  $L_{0.05,3} = Y_{(1,3)} = 48$ . When  $\gamma = 0.5$ , the calculations for  $U_{0.5,j}$ ,  $L_{0.5,j}$  and  $Q_{2,j}$  for each group are as follows:

Group 1

$n_1 = 15$ ,  $0.5 n_1 = 7.5$ ,  $g = 8$  and  $r = 0.5$ .

$$\begin{aligned} U_{0.5,1} &= \frac{1}{7.5} \left( \sum_{i=9}^{15} Y_{(i1)} + 0.5 Y_{(8,1)} \right) \\ &= \frac{1}{7.5} ((40 + 41 + \dots + 99) + (0.5)40) \\ &= 52.2667 \end{aligned}$$

$$\begin{aligned} L_{0.5,1} &= \frac{1}{7.5} \left( \sum_{i=1}^7 Y_{(i1)} + 0.5 Y_{(8,1)} \right) \\ &= \frac{1}{7.5} ((30 + 32 + \dots + 39) + (0.5)40) \\ &= 34.2667 \end{aligned}$$

$$Q_{2,1} = \frac{(99 - 30)}{(52.2667 - 34.2667)} = 3.8333$$

Group 2

$n_2 = 10$ ,  $0.5 n_2 = 5$ ,  $g = 6$  and  $r = 0$ .

$$\begin{aligned} U_{0.5,2} &= \frac{1}{5} \left( \sum_{i=6}^{10} Y_{(i2)} + (0) Y_{(5,2)} \right) \\ &= \frac{1}{5} ((42 + 43 + \dots + 64) + 0) \\ &= 50.8 \end{aligned}$$

$$\begin{aligned} L_{0.5,2} &= \frac{1}{5} \left( \sum_{i=1}^5 Y_{(i2)} + (0) Y_{(6,2)} \right) \\ &= \frac{1}{5} ((35 + 36 + \dots + 41) + 0) \\ &= 38.4 \end{aligned}$$

$$Q_{2,2} = \frac{(64 - 35)}{(50.8 - 38.4)} = 2.3387$$

Group 3

$n_3 = 10$ ,  $0.5 n_3 = 5$ ,  $g = 6$  and  $r = 0$ .

$$\begin{aligned} U_{0.5,3} &= \frac{1}{5} \left( \sum_{i=6}^{10} Y_{(i3)} + (0) Y_{(5,3)} \right) \\ &= \frac{1}{5} ((55 + 55 + \dots + 83) + 0) \\ &= 63.2 \end{aligned}$$

$$\begin{aligned} L_{0.5,3} &= \frac{1}{5} \left( \sum_{i=1}^5 Y_{(i3)} + (0) Y_{(6,3)} \right) \\ &= \frac{1}{5} ((48 + 48 + \dots + 51) + 0) \\ &= 49.8 \end{aligned}$$

$$Q_{2,3} = \frac{(83 - 48)}{(63.2 - 49.8)} = 2.6119$$

Therefore,

$$\begin{aligned} Q_2 &= \frac{(15(3.8333) + 10(2.3387) + 10(2.6119))}{(15 + 10 + 10)} \\ &= 3.0573 \end{aligned}$$

and F is classified as heavy-tailed.

Calculating Q1

Because F is classified as heavy-tailed, we have to symmetrically trim 10% of the data before calculating  $Q_1$ .

Notice that  $0.05 n_j^* < 1$  for  $j = 1, 2, 3$ .

Therefore:

$$U_{0.05,1}^* = Y_{(13,1)}^* = 52, U_{0.05,2}^* = Y_{(8,2)}^* = 56,$$

$$U_{0.05,3}^* = Y_{(8,3)}^* = 63, \text{ and } L_{0.05,1}^* = Y_{(1,1)}^* = 32,$$

$$L_{0.05,2}^* = Y_{(1,2)}^* = 36, L_{0.05,3}^* = Y_{(1,3)}^* = 48.$$

Let us calculate  $MID_j$  and  $Q_{1,j}$ , for  $j = 1, 2, 3$ .

Table 2. 10% Trimming.

Groups	Order Statistics Following 10% Symmetric Trimming	$n_j^*$
1	32 32 34 35 35 39 40 40 41 42 48 50 52	13
2	36 40 40 41 42 43 49 56	8
3	48 51 51 51 55 55 60 63	8

Group 1

$$n_1^* = 13, 0.25n_1^* = 3.25, g^* = 4, \text{ and } r^* = 0.75.$$

$$\begin{aligned} MID_1 &= \frac{1}{6.5} \left( \sum_{i=5}^9 Y_{(i)}^* + 0.75(Y_{(4,1)}^* + Y_{(10,1)}^*) \right) \\ &= \frac{1}{6.5} ((35+39+40+40+41) + (0.75)(35+42)) \\ &= 38.8846 \end{aligned}$$

$$Q_{1,1} = \frac{(52 - 38.8846)}{(38.8846 - 32)} = 1.905$$

Group 2

$$n_2^* = 8, 0.25n_2^* = 2, g^* = 3, \text{ and } r^* = 0$$

$$\begin{aligned} MID_2 &= \frac{1}{4} \left( \sum_{i=3}^6 Y_{(i2)}^* \right) \\ &= \frac{1}{4} (40 + 41 + 42 + 43) \\ &= 41.5 \end{aligned}$$

$$Q_{1,2} = \frac{(56 - 41.5)}{(41.5 - 36)} = 2.6364$$

Group 3

$$n_3^* = 8, 0.25n_3^* = 2, g^* = 3, \text{ and } r^* = 0.$$

$$\begin{aligned} MID_3 &= \frac{1}{4} \left( \sum_{i=3}^6 Y_{(i3)}^* \right) \\ &= \frac{1}{4} (51 + 51 + 55 + 55) \\ &= 53 \end{aligned}$$

$$Q_{1,3} = \frac{(63 - 53)}{(53 - 48)} = 2$$

Therefore,

$$\begin{aligned} Q_1 &= \frac{(13(1.905) + 8(2.6364) + 8(2))}{(13 + 8 + 8)} \\ &= 2.133 \end{aligned}$$

and F is classified as right skewed.

#### Discussion

As indicated in our introduction, Keselman et al. (2002) found that by first applying the Babu et al. (1999) procedure prior to testing for treatment group equality with sample symmetrically or asymmetrically determined trimmed means one could achieve excellent control over Type I errors even though data were obtained from very heterogenous distributions that were extremely nonnormal in form. Accordingly, they recommended that users adopt the Babu et al. (1999) test for symmetry.

It is also interesting to note that Babu et al. (1999) used the preliminary test for symmetry in order to determine whether groups should be compared on their symmetrically determined trimmed means, when distributions were deemed symmetric, or on their medians, when distributions were deemed asymmetric. Thus, a test for symmetry can be beneficial in many different applications.

## References

- Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal*, 413, 321-339.
- De Wet, T., & van Wyk, J. W. J. (1979). Efficiency and robustness of Hogg's adaptive trimmed means. *Communications in Statistics, Theory and Methods*, A8(2), 117-128.
- Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. *Journal of the American Statistical Association*, 70, 656-661.
- Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods*, 1(2), 288-309.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Tiku, M. L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. *Journal of Statistical Planning and Inference*, 4, 123-143.
- Tiku, M. L. (1982). Robust statistics for testing equality of means and variances. *Communications in Statistics, Theory and Methods*, 11(22), 2543-2558.
- Uthoff, V. A. (1970). An optimum test property of two well-known statistics. *Journal of the American Statistical Association*, 65, 1597-1600.
- Uthoff, V. A. (1973). The most powerful scale and location invariant test of the normal versus the double exponential. *Annals of Statistics*, 1, 170-174.
- Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, 65(1), 51-77.

## A Comparison Of The D'Agostino $S_U$ Test To The Triples Test For Testing Of Symmetry Versus Asymmetry As A Preliminary Test To Testing The Equality Of Means

Kimberly T. Perry  
Pharmacia Corporation  
Kalamazoo, Michigan

Michael R. Stoline  
Department of Statistics  
Western Michigan University

This paper evaluates the D'Agostino  $S_U$  test and the Triples test for testing symmetry versus asymmetry. These procedures are evaluated as preliminary tests in the selection of the most appropriate procedure for testing the equality of means with two independent samples under a variety of symmetric and asymmetric sampling situations.

Key words: symmetry; asymmetry; preliminary testing.

### Introduction

The purpose of this paper is to evaluate the performance of two tests, the D'Agostino  $S_U$  test and the Triples test for the testing of symmetry versus asymmetry (or skewness) as a preliminary test using two levels of significance:  $\alpha = 0.05$  and  $\alpha = 0.25$ . The results could be used to select a method for testing the equality of two means,  $H_0: \mu_1 = \mu_2$ , based on two classes of preliminary tests: (1) a test of variance homogeneity, and (2) a test of symmetry.

Procedures for the D'Agostino  $S_U$  test and the Triples test for symmetry are given below, as well as details of the four symmetric distributions and five asymmetric distributions used in the simulations. Results of a simulation study comparing the two tests for the one - sample

cases and as well as preliminary tests in two sample contexts are presented below.

### Methodology

#### Testing of Symmetry Versus Skewness

The D'Agostino test and the Triples test of symmetry are described first for a general random sample  $x_1, \dots, x_n$  from some distribution  $f(x; \mu, \sigma)$ . It is convenient to let  $\bar{x}$  denote the sample mean of  $x_1, \dots, x_n$  and to let the sample estimates of  $\beta_1^{1/2}$ , the third standardized moment, and  $\beta_2$ , the fourth standardized moment, be denoted as

$$b_1^{1/2} = m_3 / m_2^{3/2}, \quad (1)$$

$$\text{and } b_2 = m_4 / m_2^2, \quad (2)$$

$$\text{where } m_k = \sum (x_i - \bar{x})^k / n \text{ for } k = 2, 3, 4. \quad (3)$$

#### D'Agostino's Skewness Test

D'Agostino's test is a test of normality versus non-normality, which is sensitive to skewed nonnormal alternatives. A sketch of this procedure is now described.

First, compute  $b_1^{1/2}$  from the sample data. Secondly compute  $Z(b_1^{1/2})$ , where

$$Z(b_1^{1/2}) = \delta \ln(Y/a + [(Y/a)^2 + 1]^{1/2}), \quad (4a)$$

---

Kimberly T. Perry is the Director of Clinical Biostatistics, Pharmacia Corporation, Kalamazoo, Michigan. Her areas of interest are innovated clinical study designs, multiple endpoint analysis, and interim analysis. Email: kimberly.t.perry@pharmacia.com. Michael R. Stoline is Professor, Department of Statistics, Western Michigan University, Kalamazoo, Michigan. His areas of expertise are: multiple comparisons, analysis of variance, and regulatory environmental statistics. Email address: stoline@wmich.edu

$$Y = b_1^{1/2} \left[ \frac{(n+1)(n+3)}{6(n-2)} \right], \tag{4b}$$

$$W^2 = -1 + [2(\beta_2(b_1^{1/2}) - 1)]^{1/2}, \tag{4c}$$

$$\beta_2(b_1^{1/2}) = \frac{3(n^2 + 27n - 70)(n+1)}{(n-2)(n+5)(n+7)(n+9)}, \tag{4d}$$

$$\delta = 1/(\ln W)^{1/2} \text{ and } a = 2/(W^2 - 1)^{1/2}. \tag{4e}$$

The  $\alpha$ -level D'Agostino test of skewness is:

$$Z(b_1^{1/2}) > z_\alpha, \tag{5}$$

where  $z_\alpha$  is the upper  $\alpha$ -point of the standard unit normal.  $Z(b_1^{1/2})$  is approximately  $n(0, 1)$  under the null hypothesis of population normality for cases where  $n > 8$  (D'Agostino, Belanger, & D'Agostino, Jr., 1990).

Results from D'Agostino's Monte Carlo simulations for  $n < 25$  and checks with an existing table of Pearson and Hartley (1966) for  $n \geq 25$  show that the accuracy of the transformation is very good. Therefore, due to its sensitivity to skewed nonnormal alternatives, the D'Agostino test was chosen as a possible preliminary test for symmetry/skewness.

Triples Test

The Triples test is described in a paper by Randles, Fligner, Policello, and Wolfe (1980). Let  $x_1, \dots, x_n$  denote a random sample from a continuous population where  $i, j, k$  are distinct integers such that  $1 \leq i < j < k \leq n$ . The Triples test is an asymptotically distribution-free procedure which examines each triple  $(x_i, x_j, x_k)$ . If the middle observation is closer to the smaller observation than it is to the largest observation, then a "right triple" is formed (looks skewed to the right). If the middle observation is closer to the larger observation than it is to the smaller observation, then a "left triple" is formed (looks skewed to the left). The Triples test statistic is a function of the number of right triples and left triples.

The Triples test rejects  $H_0$  of symmetry if  $|T_1| > t_{n, (\alpha/2)}$  where  $t_{n, (\alpha/2)}$  is the upper  $\alpha/2$  point of a  $t$  distribution with  $n$  degrees of freedom,

$$T_1 = n^{1/2} \hat{\eta} / \hat{\sigma}_n, \tag{6a}$$

$$\hat{\eta} = \frac{\{(\text{number of right triples}) - (\text{number of left triples})\}}{\left[ \frac{3 \binom{n}{3}}{3} \right]} \tag{6b}$$

and  $\hat{\sigma}_n$  is the standard deviation of  $\hat{\eta}$ . The statistic  $\hat{\eta}$  is calculated as

$$\hat{\eta} = \binom{n}{3}^{-1} \sum_{i < j < k} f^*(x_i, x_j, x_k) \tag{7}$$

where  $f^*(x_i, x_j, x_k) = \{\text{sign}(x_i + x_j - 2x_k) + \text{sign}(x_i + x_k - 2x_j) + \text{sign}(x_j + x_k - 2x_i)\} / 3$ , and  $\text{sign}(u) = -1, 0, \text{ or } 1$  as  $u < =, \text{ or } > 0$ .

To compute  $\text{var}(\hat{\eta}) = \hat{\sigma}_n^2$ , let

$$\frac{\hat{\sigma}_n^2}{n} = \binom{n}{3}^{-1} \sum_{c=1}^3 \binom{3}{c} \binom{n-3}{3-c} \hat{\xi}_c \tag{8a}$$

$$\text{where } \hat{\xi}_c = \text{var} [f_c^*(x_1, \dots, x_c)]. \tag{8b}$$

Then  $\hat{\xi}_1 = \text{var} [f_1^*(x_1)]$ , with

$$f_1^*(x) = E [f^*(x, x_2, x_3)], \text{ yields}$$

$$\hat{\xi}_1 = \frac{1}{n} \sum_{i=1}^n (\hat{f}_1^*(x_i) - \hat{\eta})^2, \text{ where} \tag{9a}$$

$$\hat{f}_1^*(x_i) = \frac{1}{\binom{n-1}{2}} \sum_{\substack{j < k \\ j \neq i \neq k}} f^*(x_i, x_j, x_k). \tag{9b}$$

Similarly,

$$\hat{\xi}_2 = \frac{1}{\binom{n}{2}} \sum_{j < k} (\hat{f}_2^*(x_j, x_k) - \hat{\eta})^2, \text{ where} \tag{10a}$$

$$f_2^*(x_j, x_k) = \frac{1}{n-2} \sum_{\substack{i=1 \\ i \neq j \neq k \\ i \neq k}} f^*(x_i, x_j, x_k), \tag{10b}$$

$$\text{and } \hat{\xi}_3 = \frac{1}{9} - \hat{\eta}^2. \quad (11)$$

Randles, et al. (1980) compared three procedures for testing whether a univariate population is symmetric about some unspecified value compared to an immense class of asymmetric distribution alternatives. The Triples test was compared to Gupta's skewness test (Gupta, 1967) and Gupta's nonparametric procedure (Gupta, 1967). Randles et al. (1980) show that the Triples Test is superior to either competitor, even for sample sizes as small as 20, while possessing good power for detecting asymmetric alternative distributions.

Cabilio & Masaro (1996) compared their symmetry test,  $S_K$ , to several other tests of symmetry including the Triples test. The Triples test again performed well and therefore, is selected as a second possible preliminary test of symmetry/skewness.

#### Generation of Random Realizations From Six Distributions

This section contains details of how the random realizations are generated for each specified distribution among members of the normal, uniform, double exponential, logistic, lognormal, and gamma families of random variables used in the simulations. The normal, uniform, double exponential, and logistic are symmetric; the lognormal and gamma are asymmetric.

For one-sample cases, it is convenient to let  $x_1, \dots, x_n$  be a random sample of size  $n$  from  $f(x)$ . Let the sample mean and sample standard deviation be denoted as  $\bar{x}$  and  $s$ , respectively.

The IMSL random number generator RNSET, which initializes the seed, is used in all of the simulations.

#### Normal Distribution

In the case of the normal distribution, population means are set to zero,  $\mu = 0$  with unit standard deviations,  $\sigma = 1$ . The distribution  $f(x)$  is normal (0, 1). The FORTRAN function RNNOF was used to generate the normal (0, 1) random numbers.

#### Uniform Distribution

Let  $x$  be uniform ( $a, b$ ) with mean  $\mu = (a + b)/2$  and standard deviation  $\sigma = (b - a) / \sqrt{12}$ . The uniform distribution  $f(x)$  used in the simulations is a uniform (-1/2, 1/2) distribution yielding a mean  $\mu = 0$  and standard deviation  $\sigma = 1/\sqrt{12}$ .

The random numbers  $u_i$  from a uniform (0,1) distribution are first generated using the FORTRAN function RNUN. The uniform (-1/2, 1/2) random realizations are then generated using the transformation:

$$x_i = (u_i - 1/2) \quad (12)$$

#### Double Exponential Distribution

Let  $x$  have the double exponential probability density function  $f(x)$  where

$$f(x) = \frac{\exp[-|x|]}{2}, \quad -\infty < x < \infty. \quad (13)$$

The mean and variance are

$$\mu = 0 \text{ and} \quad (14)$$

$$\sigma^2 = 2. \quad (15)$$

To simulate  $x$  for this double exponential distribution, we use the following transformation:

$$x = (y_1 - y_2)/2 \quad (16)$$

where  $y_1$  and  $y_2$  are two independent chi-square random variables, each with two degrees of freedom. The two degree of freedom chi-squared random number  $y$  is generated as

$$y = -2 \ln(u) \quad (17)$$

where  $u$  is an independent random number from a uniform (0,1) distribution (see Uniform Distribution subsection).

#### Logistic Distribution

Let  $f(x)$  represent the probability density function for a logistic distribution

$$f(x) = \frac{e^x}{(1+e^x)^2} \text{ where } -\infty \leq x \leq \infty. \quad (18)$$

The mean and variance are

$$\mu = 0 \text{ and} \quad (19)$$

$$\sigma^2 = 3/\pi^2 \quad (20)$$

The random numbers  $x_i$  for this logistic distribution are generated using the transformation

$$x_i = \frac{\sqrt{3}}{\pi} \log\left(\frac{u_i}{1-u_i}\right) \quad (21)$$

where  $u_i$  is uniform (0,1).

Lognormal Distribution

The probability density function for the lognormal distribution with parameters  $a$  and  $b$  is:

$$f(x) = \ln(x; a, b) = \frac{1}{b x (2\pi)^{1/2}} \exp\left(-\frac{1}{2b^2} (\ln x - a)^2\right) \text{ for } x > 0. \quad (22)$$

The mean  $\mu$ , variance  $\sigma^2$ , and coefficient of skewness are

$$\mu = \exp\left(a + \frac{b^2}{2}\right) \quad (23)$$

$$\sigma^2 = w(w-1) \exp(2a), \text{ and} \quad (24)$$

$$\text{coefficient of skewness} = (w+2)(w-1)^{1/2} \quad (25)$$

where  $w = \exp(b^2)$ . Let  $y$  be  $n(a, b)$ , which designates a normally distributed variable with mean  $a$  and standard deviation  $b$ , then  $x = e^y$  has the lognormal probability density function  $\ln(x; a, b)$  in (22).

Three lognormal distributions are selected due to their varying degrees of skewness. In each of the three cases, the sample from the lognormal distribution  $\ln(x; a, b)$ , denoted as lognormal ( $a, b$ ), has  $a$  set to zero. The three  $b$  parameter values chosen are: (1)  $b = 0.4$ , (2)  $b = 1.0$ , and (3)  $b = 1.75$ . The coefficient of skewness for these cases are 1.3, 6.2, and 105.6, respectively. The case of  $b = 0.4$  is

denoted as slight skewness,  $b = 1.0$  as moderate skewness, and  $b = 1.75$  as heavy skewness.

The FORTRAN function RNLNL is used to create the random realizations for the  $\ln(x; a, b)$  distributions using the transformation  $x = e^y$ , where  $y$  is  $n(a, b)$  (IMSL, STAT/Library, 1989).

Gamma Distribution

The probability density function for the gamma distribution with shape parameter  $\alpha$  and scale parameter  $\beta$  is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \quad (26)$$

where  $x > 0, \alpha > 0, \beta > 0$

with mean  $\alpha\beta$ , variance  $\alpha\beta^2$  and coefficient of skewness  $2(\alpha)^{-1/2}$ .

Two gamma distributions are selected, one with shape parameter equal to 3 and unit scale parameter (denoted as G(3,1)), and the other with shape parameter equal to 2 and unit scale parameter (denoted as G(2,1)). The G(3,1) distribution is only slightly skewed (coefficient of skewness = 1.15), whereas the skewness is more pronounced in the G(2,1) distribution (coefficient of skewness = 1.41).

The gamma random realizations are generated using RNGAM (IMSL Routine) which yields random numbers with shape parameter  $\alpha$  and unit scale parameter ( $\beta = 1$ ).

## Results

### Results For Testing of Symmetry Versus Asymmetry For One Sample Cases

The robustness and the power of the D'Agostino  $S_U$  test for skewness at significance levels of  $\alpha = 0.05$  and  $0.25$ , denoted  $D(\alpha)$ , and the Triples test for symmetry at significance levels of  $\alpha = 0.05$  and  $0.25$ , denoted as  $T(\alpha)$ , are examined in this section for the one sample cases.

To assess the Type I error, the simulated null rejection rates are examined for the four symmetric distributions (normal, uniform, double exponential, and logistic). The Type I error simulated results for the two procedures are presented below. The five asymmetric distributions (lognormal (0,0.4), lognormal (0,1), lognormal (0,1.75), gamma (3,1) and gamma (2,1)) are used to

investigate the power. The simulated power results for the two tests, and discussion of the one sample results also appear below.

#### Type I Error Comparisons in One Sample Case

For the one sample cases,  $n$  random realizations are generated from each of the four symmetric distributions for each of three samples:  $n = 10, 20,$  or  $40$ . The hypothesis of symmetry is tested using the D'Agostino  $S_U$  test and the Triples test.

The two procedures are compared for control of significance level at two levels:  $\alpha = 0.05$  and  $\alpha = 0.25$  using the four symmetric distributions. A total of 10,000 simulation runs are obtained for each of the three sample sizes for each of the four symmetric distributions. Hence, twelve simulated Type I error p-values are obtained for the Triples test for the  $\alpha = 0.05$  cases, and twelve simulated p-values are also obtained for the  $\alpha = 0.25$  cases. Likewise, twelve simulated Type I error p-values are obtained for the D'Agostino  $S_U$  test for each of these two levels.

#### Significant Level Testing at 5%

For the 5% significance-level testing cases, the simulated Type I error rates (expressed as percentages) are categorized into one of the following five 5% significance level categories:

1.  $x \leq 2.5$  (extremely conservative) (27)
2.  $2.5 < x \leq 4.0$  (slightly conservative)
3.  $4.0 < x \leq 6.0$  (robust)
4.  $6.0 < x \leq 10.0$  (slightly liberal)
5.  $x > 10.0$  (extremely liberal)

The value "x" represents the percentage of rejections for testing  $H_0$ : symmetry based on the 10,000 simulations. A value "x" is obtained for each sample size and symmetric distribution combination for each procedure. Hence, twelve  $x$  values were obtained for the T(.05) cases, and twelve for the D(.05) cases.

The five 5% significance-level testing categories in (27) are labeled as robust, conservative (slightly or extremely), and liberal (slightly or extremely). These five Type I error categories are now further defined.

The outcome of the D(.05) test and the T(.05) test for a particular symmetric case is defined to be robust if the simulated null rejection rate is >

4.0 and  $\leq 6.0$ . The outcome of the D(.05) and the T(.05) test is defined to be slightly conservative if the simulated null rejection rate is  $> 2.5$  and  $\leq 4.0$ ; and extremely conservative if the simulated null rejection rate is  $\leq 2.5$ . Likewise, the test is categorized as slightly liberal if the simulated null rejection rate is  $> 6.0$  and  $\leq 10.0$ ; and extremely liberal if the simulated rejection rate is  $> 10.0$ .

The frequency and percentage of simulated Type I error rates observed in each of the five categories:  $a < x \leq b$  (given in (27)) is presented in Table 1 for the D(.05) and T(.05) tests.

#### Significance Level Testing at 25%

For the D(.25) test and the T(.25) test, the percentages of rejections (%) is tabulated for the five categories listed below:

1.  $x \leq 12.5$  (extremely conservative) (28)
2.  $12.5 < x \leq 17.5$  (slightly conservative)
3.  $17.5 < x \leq 32.5$  (robust)
4.  $32.5 < x \leq 37.5$  (slightly liberal)
5.  $x > 37.5$  (extremely liberal)

The outcome of the D(.25) test and the T(.25) test for the symmetric cases is defined to be robust if the simulated null rejection rate is  $> 17.5$  and  $\leq 32.5$ . The definitions for the conservative and liberal classifications in (28) for the D(.25) and T(.25) tests are similar to those defined in (27) for the D(.05) and T(.05) cases.

The frequency and percentage of simulated Type I error rates observed in each of the categories:  $a < x \leq b$  (given in (28)) are also presented in Table 2 for the D(.25) and T(.25) tests.

#### Discussion of Robustness for Symmetric Cases

Tables 1 and 2 show that the Triples test is more robust than the D'Agostino  $S_U$  test for symmetric cases, especially for  $\alpha = 0.25$  testing. The T(.25) test is robust in 91.7% (11 of 12) of the cases compared to 33.3% (4 of 12) of the cases for the D(.25) test. The T(.05) test is robust in 41.6% (5 of 12) of the cases compared to 25.0% (3 of 12) for the D(.05) test.



Table 1. Summary of Symmetric Distributions: Frequency of Simulated Null Rejection Rate (%) for Symmetry Versus Asymmetry Tests With Nominal 5% Level--One Sample Cases.

Test	Extremely Conservative $\leq 2.5$	Slightly Conservative $>2.5, \leq 4.0$	Robust $>4.0, \leq 6.0$	Slightly Liberal $>6.0, \leq 10$	Extremely Liberal $>10$
D(.05)	3 (25.0%)	0 (0.0%)	3 (25.0%)	0 (0.0%)	6 (50.0%)
T(.05)	3 (25.0%)	2 (16.7%)	5 (41.6%)	2 (16.7%)	0 (0.0%)

*Note:* Table 1 results are based on the four symmetric distributions (normal, uniform, double exponential, and logistic) and three sample sizes ( $n = 10, 20$  and  $40$ ).

Table 2. Summary of Symmetric Distributions: Frequency of Simulated Null Rejection Rate (%) for Symmetry Versus Asymmetry Tests With Nominal 25% Level--One Sample Cases.

Test	Extremely Conservative $\leq 12.5$	Slightly Conservative $>12.5, \leq 17.5$	Robust $>17.5, \leq 32.5$	Slightly Liberal $>32.5, \leq 37.5$	Extremely Liberal $>37.5$
D(.25)	2 (16.7%)	1 (8.3%)	4 (33.3%)	1 (8.3%)	4 (33.3%)
T(.25)	0 (0.0%)	1 (8.3%)	11 (91.7%)	0 (0.0%)	0 (0.0%)

*Note:* Table 2 results are based on the four symmetric distributions (normal, uniform, double exponential, and logistic) and three sample sizes ( $n = 10, 20$  and  $40$ ).

Tables 1 and 2 also show that the D'Agostino  $S_U$  test is appreciably more liberal than the Triples test for symmetric cases. The D(.05) test is observed to be liberal in 50.0% (6 of 12) of the cases compared to 16.7% (2 of 12) for the T(.05) test. Also, the D(.25) test is observed to be liberal in 41.6% (5 of 12) of the cases compared to 0.0% (0 of 12) of the T(.25) cases.

On the basis of the results presented in Tables 1 and 2, it is concluded that the Triples Test is superior to the D'Agostino  $S_U$  test for controlling Type I error. It is also concluded that the D'Agostino  $S_U$  test does not control the Type I error rate for symmetric cases since it fails to maintain the Type I error rate at or below the stated level of significance.

#### Results of Power Analysis in One Sample Cases

The results of a power comparison of the D'Agostino  $S_U$  test and the Triples test is now reported. A total of 10,000 simulation runs are obtained for each of the three sample sizes  $n = 10, 20,$  and  $40$  for each of the five asymmetric

distributions. Hence, fifteen simulated power  $p$ -values are obtained for the Triples test for each of the T(.05), T(.25), and D(.05), and D(.25) cases.

#### Definition of Power Categories

The results of the simulation for the five asymmetric distributions are combined in Table 3 over all sample sizes for the four power categories defined below:

1.  $x \leq 50.0$  (low power) (29)
2.  $50.0 < x \leq 75.0$  (moderate power)
3.  $75.0 < x \leq 90.0$  (high power)
4.  $x > 90.0$  (extremely high power)

The value " $x$ " represents the power to detect asymmetry based on 10,000 simulations for each sample size configuration. Each entry in Table 3 denotes both the frequency and percentage at which  $a < x \leq b$  occurs, as in Table 1.

The four power categories in (29) are conveniently labeled in order of increasing power: low power (power  $< 50\%$ ), moderate power ( $50\% < \text{power} \leq 75\%$ ), high power ( $75\% < \text{power} \leq 90\%$ ), and extremely high power (power  $> 90\%$ ).

Table 3. Summary of Asymmetric Distributions: Frequency of Simulated Power Rates (%) for Symmetry Versus Asymmetry Tests With Nominal 5% and 25% Levels, One Sample Cases.

Test	Low Power $\leq 50.0$	Moderate Power $>50.0, \leq 75.0$	High Power $>75.0, \leq 90.0$	Extremely High Power $>90.0$
Nominal 5% Level				
D(.05)	6 (40.0%)	3 (20.0%)	3 (20.0%)	3 (20.0%)
T(.05)	7 (46.7%)	3 (20.0%)	2 (13.3%)	3 (20.0%)
Nominal 25% Level				
D(.25)	2 (13.3%)	3 (20.0%)	3 (20.0%)	7 (46.7%)
T(.25)	3 (20.0%)	4 (26.7%)	2 (13.3%)	6 (40.0%)

Note: Table 3 results are based on the asymmetric distributions [lognormal ( 0, 0.40), lognormal ( 0, 1.0), lognormal ( 0, 1.75), G(3,1), and G(2,1)] and three sample sizes (n = 10, 20 and 40).

These four power categories are used in Table 3 for both 5% and 25% results.

Discussion of Power for Asymmetric Cases

Table 3 shows that both the T(.05) and D(.05) tests lack power. The power is  $\leq 0.75$  for 60% of the cases when using the D(.05) test, and is  $\leq 0.75$  for 66.7% of the cases when using the T(.05) test. The D(.05) test is generally more powerful than the T(.05) test for asymmetric cases.

The D(.25) test tends to be somewhat more powerful than the T(.25) test. The power is  $> .90$  for approximately 47% of the cases when using the D(.25) test compared to 40% of the cases when using the T(.25) test. In addition, the power is  $\leq 0.50$  for 20% of the cases when using the T(.25) test compared to approximately 13% when using the D(.25) test.

It is concluded that the D'Agostino  $S_U$  test is somewhat more powerful than the Triples test for detecting asymmetric distributions.

Discussion of One Sample Simulation Results

Table 4 contains summary statistics describing the mean, standard deviation (denoted as  $s$ ), minimum, and maximum of the four sets of twelve simulated p-values obtained by using the D(.05), T(.05), D(.25), and T(.25) procedures for the symmetric cases. The symmetric case

summary statistics can be used to characterize the Type I error properties of these test procedures.

The symmetric mean p-value is denoted as  $\bar{p}_s$  in Table 4.

Table 5 also contains the corresponding summary statistics of the four sets of fifteen simulated p-values obtained by the same four test procedures for the asymmetric cases. The asymmetric case summary statistics can be used to characterize the power properties of these procedures. The asymmetric mean p-value is denoted as  $\bar{p}_a$  in Table 5.

For the symmetric cases summarized in Table 4, the average Type I error rates for the T(.05) and T(.25) procedures are  $\bar{p}_s = 4.1\%$  and  $\bar{p}_s = 21.5\%$ , respectively, compared to  $\bar{p}_s = 11.2\%$  and  $\bar{p}_s = 31.0\%$  for the D(.05) and D(.25) procedures, respectively. The average Type I error rates for the Triples test are observed to be closer to the stated significance levels of 5% and 25% than are those for the D'Agostino  $S_U$  test.

For the symmetric cases summarized in Table 4, the standard deviations  $s$  and ranges of the p-values for the T(.05) and the T(.25) procedures are appreciably smaller than the comparable standard deviations and ranges for the D(.05) and the D(.25) procedures.

Table 4. Descriptive Statistics of the Simulated p-values for Four Test Procedures: D(.05), T(.05), D(.25), and T(.25) for Symmetric Cases (Summary statistics displayed as percentages)

Type I Error Summary Statistics	Significance level 5%		Significance level 25%	
	D(.05)	T(.05)	D(.25)	T(.25)
$\overline{p_a}$	11.2	4.1	31.0	21.5
$s$	10.5	1.6	16.0	2.6
minimum	0.2	1.6	8.0	16.3
maximum	33.2	6.3	58.0	25.0
n	12	12	12	12

Table 5 contains the corresponding summary statistics of the four sets of fifteen simulated p-values obtained by the same four test procedures for the asymmetric cases. The asymmetric case summary statistics can be used to characterize the power properties of these procedures. The asymmetric mean p-value is denoted as  $\overline{p_a}$  in Table 5.

Table 5. Descriptive Statistics of the Simulated p-values for Four Test Procedures: D(.05), T(.05), D(.25), and T(.25) for Asymmetric Cases (Summary statistics displayed as percentages).

Power Summary Statistics	Significance level 5%		Significance level 25%	
	D(.05)	T(.05)	D(.25)	T(.25)
$\overline{p_a}$	60.4	52.6	80.0	75.2
$s$	30.0	34.2	20.1	24.0
minimum	16.8	5.7	44.3	33.0
maximum	100.0	100.0	100.0	100.0
n	15	15	15	15

### Summary

For symmetric cases summarized in Tables 1, 2, and 4, it is concluded that the Triples test is superior to the D'Agostino  $S_U$  test for the control of Type I error. The Triples test tends to hold to the stated level of significance. The D'Agostino  $S_U$  test does not hold to the stated level of significance and often tends to be liberal.

For the asymmetric cases summarized in Table 5, the average powers of the T(.05) and the

T(.25) procedures are  $\overline{p_a} = 52.6\%$  and  $\overline{p_a} = 75.2\%$ , respectively, compared to  $\overline{p_a} = 60.4\%$  and  $\overline{p_a} = 80.0\%$ , respectively for the D(.05) the D(.25) procedures. The D'Agostino  $S_U$  test is observed to be slightly more powerful than the corresponding Triples test. The D'Agostino  $S_U$  test may be more powerful for asymmetric alternatives because the D'Agostino  $S_U$  test tends to be liberal with respect to Type I error control.

### Testing Symmetry Versus Asymmetry In Preliminary Testing For Two Sample Cases

A purpose of this study is to select a preliminary test of testing symmetry versus asymmetry, and using the preliminary test to select the most appropriate method for testing the equality of two independent means  $H_0: \mu_1 = \mu_2$ . A two sample  $t$  procedure is commonly used if the underlying distributions are symmetric, and a Mann-Whitney-Wilcoxon (MWW) procedure may be more appropriate if the underlying distributions are asymmetric. The decision to use the  $t$  or the MWW procedure is often based on the personal preference of the investigator, or an examination of descriptive and graphical comparative statistics between the two samples.

Little evidence exists in the statistical literature of the use of tests of symmetry versus asymmetry as a preliminary test to select the  $t$  or MWW methods prior to testing  $H_0: \mu_1 = \mu_2$ . In these situations, the  $t$  procedure would be used if the preliminary test for skewness is non-significant; otherwise, the MWW procedure is used.

### Two Sample Preliminary Testing Strategies

Assume there are two independent samples of sizes  $n_1$  and  $n_2$  from two distributions  $f_1(x_1; \mu_1, \sigma_1)$  and  $f_2(x_2; \mu_2, \sigma_2)$ , respectively. Let us assume that the same skewness test is applied to the data from the two samples separately where the same significance level  $\alpha$  is used for both tests.

Two preliminary testing protocols are defined. One utilizes the MWW test of  $H_0: \mu_1 = \mu_2$  if at least one (ALO) of the two preliminary skewness tests is significant. The other utilizes the MWW test if both (BOTH) preliminary tests are significant. There two preliminary testing

strategies are conveniently labeled: ALO and BOTH.

Selection of a Preliminary Testing Strategy

The one-sample simulation results summarized in Tables 4 and 5 are used to select a preliminary testing method between the BOTH and ALO protocols. For this purpose, it is convenient to utilize the average p-values:  $\bar{p}_s$  and  $\bar{p}_a$  p-values of the twelve symmetric and fifteen asymmetric distributions, respectively, summarized in Tables 4 and 5 for the D(.05), T(.05), D(.25), and T(.25) one-sample skewness test procedures.

Assuming symmetry (SYM) is true, the probability of correct selection of the  $t$  method for testing  $H_0: \mu_1 = \mu_2$  is approximately given as:

$$1 - \bar{p}_s^2 \text{ for the BOTH method, and } (30a)$$

$$(1 - \bar{p}_s)^2 \text{ for the ALO method. } (30b)$$

Assuming asymmetry (ASY) is true, the probability of correct selection of the MWW method for testing  $H_0: \mu_1 = \mu_2$  is approximately given as:

$$\bar{p}_a^2 \text{ for the BOTH method, and } (31a)$$

$$1 - (1 - \bar{p}_a)^2 \text{ for the ALO method. } (31b)$$

Table 6 contains the probabilities of correct preliminary test selection of the  $t$  or MWW method for testing  $H_0: \mu_1 = \mu_2$  depending on whether the underlying distribution is symmetric (SYM) or asymmetry (ASY), and whether the BOTH or ALO preliminary test strategy is used.

For SYM cases, the BOTH method has the higher probabilities of correct selection of the  $t$  test since:  $1 - \bar{p}_s^2 > (1 - \bar{p}_s)^2$ . Whereas for ASY cases, the ALO method has the higher probabilities of correct selection of the MWW test since:  $1 - (1 - \bar{p}_a)^2 > \bar{p}_a^2$ .

Table 6. Probabilities of Correct Preliminary Test Selection of the Method to Test  $H_0: \mu_1 = \mu_2$

Correct Selection Probability	Preliminary Test Protocol	Underlying Distribution	Correct Methods
$1 - \bar{p}_s^2$	BOTH	SYM	$t$
$(1 - \bar{p}_s)^2$	ALO	SYM	$t$
$\bar{p}_a^2$	BOTH	ASY	MWW
$1 - (1 - \bar{p}_a)^2$	ALO	ASY	MWW

Table 7 contains the estimated probabilities of correct preliminary test method selection described in Table 6 for the various methods. The estimated probabilities in Table 7 are calculated utilizing the average  $\bar{p}_s$  and  $\bar{p}_a$  values tabled in Tables 4 and 5.

Table 7. Estimated Preliminary Test Probabilities of Correct Selection of the Method to Test  $H_0: \mu_1 = \mu_2$

Method	$\bar{p}_s$	$\bar{p}_a$	BOTH		ALO	
			SYM	ASY	SYM	ASY
D(.05)	.112	.604	.987	.365	.789	.843
T(.05)	.041	.526	.998	.277	.920	.775
D(.25)	.310	.800	.904	.640	.476	.960
T(.25)	.215	.752	.954	.556	.616	.938

Discussion

Preliminary testing methods are recommended that maximize the Table 7 probabilities of correct selection for the SYM and ASY cases. Using this criterion, the BOTH method is preferred for correct  $t$  test selection for SYM cases, and the ALO method is preferred for correct MWW test selection for ASY cases. Also, the 5% significance level is preferred for SYM cases, and the 25% level is preferred for ASY cases. Furthermore, the Triples tests are preferred for SYM cases, and the D'Agostino  $S_U$  tests are preferred for ASY cases.

How then can a single preliminary testing strategy be selected if different strategies, significance levels, and methods are preferred for SYM versus ASY cases?

To resolve this question another preliminary test comparison criterion is introduced.

Preliminary testing methods are recommended that tend to provide equal or nearly equal probabilities of correct method selection for both SYM and ASY cases. Using this criterion with the results in Table 7, two methods are recommended for preliminary test usage. These are the T(.05) and D(.05) procedures, where both use the ALO method.

The probabilities of correct method selection are 0.920 for SYM cases and 0.775 for ASY cases using the T(.05) ALO method. The corresponding probabilities are 0.789 and 0.843, respectively, for the D(.05) ALO method. No other procedures in Table 7 have this high degree of balance between the equality of probabilities of correct model selection for typical SYM and ASY cases. The T(.05) method is preferred if more emphasis is needed for correct method selection for SYM cases, whereas, the D(.05) method is preferred if more emphasis is needed for correct method selection for ASY cases.

### Conclusion

#### One Sample Symmetry Versus Asymmetry Tests

The one sample Triples test is superior to the D'Agostino  $S_U$  test for the control of Type I error for symmetric cases, whereas, the one sample D'Agostino  $S_U$  test is slightly more powerful than the Triples tests for asymmetric alternatives.

#### Preliminary Test Of Symmetry Versus Asymmetry Prior To A Test Of Equality Of Means

The Triples test using a 5% level of significance is preferred if more emphasis is needed for correct method selection for symmetric cases, whereas, the D'Agostino  $S_U$  test using a 5% level of significance level is preferred if more emphasis is needed for correct method selection for asymmetric cases.

### Recommendations

A simulation study examining the characteristics of the use of a preliminary test of skewness versus asymmetry prior to testing  $H_0: \mu_1 = \mu_2$  would be of interest. On the basis of the analyses reported here, the Triples test or the D'Agostino  $S_U$  test with a 5% level of significance is recommended over the Triples test or the D'Agostino  $S_U$  test with a 25% level of significance as a preliminary test of skewness versus asymmetry prior to testing  $H_0: \mu_1 = \mu_2$ .

### References

- Cabilio, P. & Masaro, J. (1996). A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics*, 24(3), 349-361.
- D'Agostino, R. B., Belanger, A. & D'Agostino Jr., R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44 (4), 316-321.
- Gupta, M. K. (1967). An asymptotically nonparametric test of symmetry. *Annals of Mathematical Statistics*, 38, 849-866.
- IMSL. (1989, December). *Math/Library User's Manual (Version 1.1)*. Houston, Texas: Author.
- IMSL. (1989, January). *Stat/Library User's Manual (Version 1.1)*. Houston, Texas: Author.
- Pearson, E. S. & Hartley, H. O. (1966). *Biometrika Tables for Statisticians. Vol. I*, 3rd edition. Cambridge University Press.
- Randles, R. H., Fligner, M. A., Policello II, G. E., & Wolfe, D. A. (1980, March). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, 75 (369), 168-172.

## On the Estimation of Binomial Success Probability With Zero Occurrence in Sample

Mehdi Razzaghi  
 Mathematics, Computer Science, & Statistics  
 Bloomsburg University

The problem of estimating the probability of a rare event when the sample shows no incidence of the event is considered. Several methodologies based on various statistical techniques are described and their relative performances are investigated. A decision theoretic approach for estimation of response probability when the sample contains zero responses is examined in depth. The properties of each method are discussed and an example from teratology is used to provide illustration and to demonstrate the results.

**Key words:** Binomial distribution, response probability estimation.

### Introduction

There are many instances in practice that an estimate of the probability of occurrence of a rare event is desired. Because of the low probability of the event, however, the experimental data may conceivably indicate no occurrence of that event. For example, in cancer risk estimation with laboratory animals, often at low doses, data may exhibit no animals with tumors, even though there is a nonzero probability of response at that dose. More specifically, suppose that  $X$  is the number of occurrences of an event in a sample of  $n$  independent and identical Bernoulli trials. Then  $X$  has a binomial distribution with

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad x = 0, 1, \dots, n \quad (1)$$

where  $p$  is the probability of occurrence in each trial. It is well known that the maximum

likelihood estimate of  $p$  is  $x/n$ . But when  $x = 0$ , this estimate is often unrealistic and alternative methods should be utilized to estimate  $p$ . Observation of zero occurrence in a sample is not uncommon in practice. Table 1 provides numerical values of the probability of zero successes in binomial experiments for different sample sizes.

Table 1. Probability of zero response for varying sample sizes and different true response probabilities.

$p \backslash n$	0.01	0.02	0.05	0.07	0.10	0.15	0.20
1	0.990	0.980	0.950	0.930	0.900	0.850	0.800
2	0.980	0.960	0.902	0.865	0.810	0.722	0.640
4	0.961	0.922	0.814	0.748	0.656	0.522	0.410
10	0.904	0.817	0.599	0.484	0.349	0.197	0.107
20	0.818	0.668	0.358	0.234	0.122	0.039	0.011
30	0.740	0.545	0.215	0.113	0.423	0.008	0.001

Note that even when  $p$  is as high as 0.05 and the sample is as high as twenty, there is still a 36% chance of no response in the data. Bailey (1997) considered the problem of estimating  $p$  when the sample has no occurrence and proposed a method currently used in risk analysis of energetic initiation in the explosive testing field. This estimator is given by

$$\hat{p} = 1 - (0.5)^{1/n} \quad (2)$$

Professor Mehdi Razzaghi's area of interest is environmental statistics with applications of statistical modeling and risk assessment in toxicological experiments. Address: Mathematics, Computer Science, & Statistics, Bloomsburg University, Bloomsburg, PA 17815. E-mail: [razzaghi@bloomu.edu](mailto:razzaghi@bloomu.edu). The author is grateful to the Editor and the anonymous referees for their helpful comments.

which is obtained by setting the probability of observing  $n$  failures equal to 0.5 and solving for  $p$ . Bailey noted that this estimator is nearly identical to the median of the Bayesian posterior distribution for  $p$ , derived with respect to a uniform distribution using the absolute error loss (AEL) function.

The problem of Bayesian estimation of  $p$  with respect to the more general class of a conjugate beta prior distribution but using the squared error loss (SEL) was considered by Basu et al. (1996). By comparing (2) with a few other estimates, Bailey (1997) concluded that  $\hat{p}$  performs relatively well in practice and can be used in certain circumstances. It is also worth noting that because the upper  $100(1 - \alpha)\%$  confidence limit for  $p$  is (see Bickel & Doksum, 2001) given by

$$u = 1 - \alpha^{1/n}$$

then (2) can be interpreted as the median of the sampling distribution of the random variable  $X/n$ . Moreover, as mentioned in Louis (1981),  $u$  may be thought of as the proportion of the number of successes in a future experiment of the same size and it is the upper  $100(1 - \alpha)\%$  Bayesian prediction interval based on a uniform prior distribution.

In this paper, the problem of point estimation of  $p$  when a sample shows no occurrence is considered from a more general viewpoint. Several potential estimates based on statistical methods in addition to those suggested in Bailey (1997) and Basu et al. (1996) will be proposed and their properties will be discussed. Next, I review the Bayesian approach and consider the use of other loss functions, and then discuss the properties of an estimate derived from information theory. The next section is devoted to the discussion of a decision theoretic approach for estimating  $p$ , and the use of minimax estimation of  $p$  is considered. In the final section of this article, I give an example from teratology to provide further illustration of the results.

### Bayesian Estimation

It is well known that when the prior distribution of  $p$  belongs to the family of a beta distribution  $\beta(a, b)$ ,

$$g(p) = \frac{1}{B(a, b)} p^{a-1} (1-p)^{b-1} \quad a, b > 0, 0 < p < 1 \quad (3)$$

where

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

then the posterior distribution of  $p$  belongs to the beta family  $\beta(a+x, b+n-x)$  and the Bayes estimate  $p^*$  of  $p$  based on the SEL function  $L(p, p^*) = (p - p^*)^2$ , is given by (Basu et al., 1996)

$$p^* = \frac{(a+x)}{(a+b+n)} \quad (4)$$

Thus, if  $x = 0$ , then the Bayes estimator for a zero occurrence is

$$p^* = \frac{a}{a+b+n} \quad (5)$$

and in particular if  $a = b = 1$ , then the Bayes estimator under a uniform prior is derived. Also, when Jeffreys' non-informative prior, for which  $a = b = 0.5$  is used, then the Bayes estimator of no response is given by

$$p_{ni}^* = \frac{1}{2(n+1)} \quad (6)$$

Basu et al. (1996) compared (5) and (6) with the classical approach based on upper confidence limits and conclude that the Bayes estimate under an informative prior is best. Both estimates (5) and (6), however, are derived using the SEL function which is but one of several possible loss functions that may be used to derive the Bayes estimate of  $p$ . In practice, there are many instances that other functions may be preferred.

Actually the SEL is a special case of a larger class of weighted quadratic loss functions

$$L(p, p^*) = w(p)(p - p^*)^2 \quad (7)$$

where  $w(p) \geq 0$  is an appropriate weight function. For the class (7) the posterior expected loss is minimized when

$$p^* = \frac{E(pw(p))}{E(w(p))} \tag{8}$$

where the expectation is with respect to the posterior distribution of  $p$ . In particular if  $w(p)$  is of the form

$$w(p) = p^\alpha (1-p)^\beta \tag{9}$$

for some  $\alpha$  and  $\beta$ , then from (8)

$$p^* = \frac{E(p^{\alpha+1} (1-p)^\beta)}{E(p^\alpha (1-p)^\beta)} \tag{10}$$

which for the family of beta prior, yields

$$p^* = \frac{a + \alpha + x}{a + b + n + \alpha + \beta} \tag{11}$$

Now, if  $\alpha = \beta = 0$ , then (4) is obtained as a special case of this larger class of estimates. Another special case, and possibly more appropriate for the purpose of risk assessment, in (11) is when  $\alpha = \beta = -1$ , corresponding to the scaled square error loss (SSEL) function

$$L(p, p^*) = \frac{(p - p^*)^2}{p(1-p)}$$

In this case, however, it is easy to see that when  $x = 0$ , and  $a = 1$ , then  $p^*$  is the only estimate which produces an infinite posterior expected loss. Hence, when there is no occurrence in the sample the SSEL function does not produce a useful solution. Indeed, when  $x = 0$ , the SSEL function produces a negative estimate of  $p$  for  $a < 1$ . Note also from (11) that in this case the Bayes estimate with respect to a uniform distribution is identical to the maximum likelihood estimate.

Aside from the class of squared error loss functions, a class of functions often used in

Bayesian estimation is the absolute error loss (AEL) given by

$$L(p, p^*) = |p - p^*|,$$

for which the Bayes estimate is the median of the posterior distribution. Hence for the family of beta prior (3), when  $x = 0$ , we seek  $p_1^*$  such that

$$I_{p_1^*}(a, b+n) = \int_0^{p_1^*} \frac{1}{B(a, b+n)} p^{a-1} (1-p)^{b+n-1} dp = 0.5 \tag{12}$$

which for given values of  $a$  and  $b$  can be evaluated using tables of incomplete beta functions (e.g. Pearson & Hartley, 1956) or any standard numerical technique. Specifically, if  $a = b = 1$ , then (12) yields

$$p_1^* = 1 - (0.5)^{1/(n+1)} \tag{13}$$

which, as noted earlier, is for large  $n$  approximately equal to the Bailey (1997) estimate. Also, when Jeffrey's non-informative prior ( $a = b = 0.5$ ) is used, an approximation to the solution of (12) may be obtained by using a procedure described in Johnson and Kotz (1995) regarding the approximations to the beta function ratio.

Accordingly, if  $p_{1, n_i}^*$  denotes the solution of (12) for  $a = b = 0.5$ , then an approximate value of  $p_{1, n_i}^*$  can be obtained as the solution of

$$n + \frac{1}{6} - \left(n + \frac{7}{3}\right)(1-x) + \frac{1}{5} \frac{2x}{2n+1} + \frac{x-1/2}{n+1} = 0 \tag{14}$$

where the error of approximation is generally below .001.

Another choice of a loss function for Bayesian estimation is the so-called zero-one loss defined as

$$L(p, p^*) = \begin{cases} 0 & \text{if } |p - p^*| \leq \varepsilon \\ 1 & \text{if } |p - p^*| > \varepsilon \end{cases}$$



which amounts to no loss if the estimate  $p^*$  is within a distance  $\varepsilon$  from  $p$ . For this loss function, the expected posterior is given by

$$P(|p - p^*| > \varepsilon | x) = 1 - P(|p - p^*| \leq \varepsilon | x).$$

Consequently, if a modal interval of length  $2\varepsilon$  is defined as an interval with center at the mode of the distribution, then as  $\varepsilon \rightarrow 0$ , the Bayes estimate with respect to the zero-one loss approaches the mode of the posterior distribution, provided that a mode exists. This in turn implies that the Bayes estimate in this case becomes the maximum likelihood estimate.

#### Maximum Information Estimation

Good (1965) and Typlados and Brimley (1962) showed that Shannon's information content of the observation  $x$  from the binomial distribution (1) is given by

$$I(p) = -p \ln(p) - (1-p) \ln(1-p) + \ln \left[ \binom{n}{x} p^x (1-p)^{n-x} \right] \quad (15)$$

By maximizing  $I(p)$ , one obtains the maximum information (MIE) estimate  $p_{MIE}$  of  $p$  as the solution of the equation

$$\ln \left( \frac{p}{1-p} \right) = \frac{x}{p} - \frac{n-x}{1-p} \quad (16)$$

In particular when  $x = 0$ , the MIE of  $p$  is the solution of

$$\sqrt[n]{\frac{p}{1-p}} = \exp \left( -\frac{1}{1-p} \right). \quad (17)$$

Chew (1971) pointed out that for  $n > 7$ , the solution of (17) is up to 3 decimals equal to zero and, once again, it is seen that this method fails to produce a reasonable estimate for  $p$ .

#### Minimax Estimation

The minimax criterion stems from the general theory of two-person zero-sum games of von Neuman and Morgenstern (1944). Loosely,

instead of averaging the risk as in Bayesian estimation, one looks at the least favorable scenario for each decision, that is the worst possible risk for that decision, and chooses a decision which gives the least value of the worst risk. Thus, the minimax rule minimizes the maximum risk. Although the methodology ignores all references to prior knowledge, but in the absence of any information regarding  $p$ , the minimax estimator provides a Bayesian estimate without knowing the prior distribution. As pointed out by Cox and Hinkley (1974), the minimax rule is defensible when the risk is small, since it ensures that, whatever the true parameter value, the expected loss is small. Although there may be an apparently better rule, any improvement can only be small and may carry with it the danger of a seriously bad performance for some values of the parameter.

Now, for the binomial parameter  $p$  in (1), it can be shown that the minimax decision rule, based on the SEL function, is given by (Bickel and Doksum, 2001)

$$\tilde{p} = \frac{x + \sqrt{n}/2}{n + \sqrt{n}} \quad (18)$$

with variance bounded by

$$v = [2(1 + \sqrt{n})]^{-2} \quad (19)$$

The minimax estimator (18) is Bayes with respect to a beta prior with parameters

$\sqrt{n}/2$  and  $\sqrt{n}/2$ . If  $x = 0$ , then from (18),

$$\tilde{p} = [2(1 + \sqrt{n})]^{-1} \quad (20)$$

which can be used to estimate the probability of a rare event. In order to compare the minimax estimator given in (20) with those considered in Bailey (1997),  $\tilde{p}$  was evaluated for several values of  $n$ . Table 2 presents these numerical values, where for comparison, the values of  $\hat{p}$  in (2), the estimator suggested by Bailey and the Bayes estimator  $p_{ni}^*$  based on a noninformative prior given in (6) are also included. As the sample size

increases, the minimax method appears to produce numerically larger point estimates.

Table 2. Numerical values of minimax ( $\tilde{p}$ ), Bayes ( $p_{ni}^*$ ) and Bailey ( $\hat{p}$ ) estimator.

n	1	2	4	10	20	30	40
$\tilde{p}$	.250	.207	.167	.120	.091	.077	.062
$p_{ni}^*$	.250	.167	.100	.045	.024	.016	.010
$\hat{p}$	.500	.293	.159	.067	.034	.023	.014

Because the binomial distribution  $E(X) = np$ , it is clear from (4) and (18) that

$$E(p_{ni}^*) = \frac{2np + 1}{2(n + 1)} \tag{21}$$

and

$$E(\tilde{p}) = \frac{2\sqrt{np} + 1}{2(\sqrt{n} + 1)} \tag{22}$$

Table 3 provides the numerical values of (21) and (22) for selected values of  $n$  and  $p$  where for completeness we also include a crude estimate of  $E(\hat{p})$ , computed by using (2) for  $x = 0$ .

p	.01			.05			.10		
	$\tilde{p}$	$p_{ni}^*$	$\hat{p}$	$\tilde{p}$	$p_{ni}^*$	$\hat{p}$	$\tilde{p}$	$p_{ni}^*$	$\hat{p}$
4	0.173	0.108	0.169	0.200	0.140	0.209	0.233	0.180	0.259
10	0.128	0.054	0.077	0.158	0.091	0.117	0.196	0.136	0.167
20	0.099	0.033	0.044	0.132	0.071	0.084	0.173	0.119	0.134
30	0.086	0.026	0.033	0.119	0.064	0.073	0.162	0.113	0.123
40	0.077	0.022	0.027	0.111	0.061	0.067	0.155	0.110	0.117
50	0.071	0.020	0.023	0.105	0.059	0.064	0.149	0.108	0.114

Table 3. Expected values of minimax ( $\tilde{p}$ ), Bayes ( $p_{ni}^*$ ) and Bailey ( $\hat{p}$ ) estimators for varying sample sizes and for different true response probabilities.

Example

Kochhar et al. (1992) describes an experiment to examine the developmental toxicity of two retinoylamino acids, RG and RL in IRC mice and compare them with other retinamides. One of the observed effects was the incidence of cleft palate in the viable fetuses. Table 4 presents the percentage of fetuses with cleft palate for different doses together with the number of implants per dose group as a result of maternal exposure to retinoic acid (RA).

Table 4. Incidence of cleft palate in offspring of mice exposed to retinoic acid (RA). Source: Kochhar et al. (1992).

Dose mg/kg	0	5	10	25	100
Number of Implants	152	98	78	86	164
% with Cleft Palate	1	0	13	33	82

It is observed that even though there was 1% response rate in the control group, there was no occurrence of cleft palate in the 5 mg/kg dose group. The incidence rate in other dose groups showed a statistically significant difference from the control group. For risk assessment purposes, in practice one would fit a suitable dose-response model to these data and extrapolate to low exposure levels to obtain an upper confidence limit for the risk at a fixed low dose.

The model can equivalently be used to obtain a benchmark dose, which is the lower confidence limit for dose corresponding to a given low negligible level of risk. However, because of no incidence at the lowest non-zero dose level, one might erroneously consider fitting a non-monotonic dose-response function.

That is, the analysis might lead to the conclusion that the chemical has a hormetic effect, i.e. it is low dose stimulative and high dose inhibitive. For a discussion on the concept of

chemical hormesis we refer to Calabrese and Baldwin (2000). However, as shown in Razzaghi and Loomis (2001), in developmental toxicology, more than a single replication of an experiment must be considered before a chemical can be declared as being hormetic. For the present data, therefore, in order to fit a monotonic dose-response function, one might consider replacing the observed incidence of zero by an estimate of it. In such a situation, it would seem unreasonable to estimate the probability of response in the 5 mg/kg dose group as 0, as given by the maximum likelihood method. In this case, because  $n = 98$ , from (2), (6), (14) and (20),

$\hat{p} = .007$ ,  $p_{ni}^* = .005$ ,  $p_{1,ni}^* = .021$ ,  $\tilde{p} = 0.046$  are four different point estimates for the probability of response at the first nonzero dose level.

In order to further investigate the properties of these estimates, a probit model was used to fit the response probability  $p$  as a function of the natural logarithm of dose, i.e.

$$p = \Phi(a + b \log d) \quad (23)$$

Using PROC PROBIT in SAS (1996), it was found that the maximum likelihood estimates of the model parameters are  $\hat{a} = 03.601$  and  $\hat{b} = 0.987$ . Using these parameter estimates, it is found that the point estimate of  $p$  when  $d = 5$  mg/kg is .022. Furthermore, the standard deviation of  $\hat{a} + \hat{b} \log 5$  is 0.163. Based on these quantities, if the 95% confidence interval is evaluated for the predicted proportion, one finds that this range is (.010, .046). Interestingly, although the minimax estimator  $\tilde{p}$  is equal to the upper bound in this range, both the Bailey estimator  $\hat{p}$  and the Bayesian estimator  $p_{ni}^*$  are outside this range and far too small to be plausible. Therefore, in this instance,  $p_{1,ni}^*$  and the minimax procedure appear to produce more realistic estimates of  $p$  compared to other methods.

### Discussion

Lack of occurrence of rare events in biological and physical experiments is not uncommon. In such situations, the maximum likelihood estimate

becomes unusable and one needs to resort to alternative statistical methods. Here, I have considered this problem and investigated the use of several other statistical techniques and the minimax estimator.

It is immediately noted from (2) that for the Bailey estimator,  $\hat{p} = 0\left(\frac{1}{n}\right)$ . This property also holds for the Bayesian estimator considered by Basu et al. (1996). However, for the minimax estimator, from (18)  $\tilde{p} = 0\left(\frac{1}{\sqrt{n}}\right)$ . This means that

for relatively large values of  $n$ , both  $\hat{p}$  and the Bayes estimate lead to numerically smaller values than the minimax estimator. Actually, it can be shown (Roussas, 1997) that the Bayes estimate for the family of beta prior and SEL has the same asymptotic distribution as the maximum likelihood estimate for arbitrary fixed values of  $\alpha$  and  $\beta$ , while the asymptotic distribution of  $\sqrt{n}(\tilde{p} - p)$  is normal with mean  $\frac{1}{2} - p$  and variance  $p(1-p)$ .

Thus, I can say that the minimax estimator is comparatively more conservative.

However, as discussed by Carlin and Louis (1996), although informative priors enable more precise estimation, extreme care must be taken in their use because they also carry the risk of disastrous performance when their informative content is in error. Although using a non-informative prior leads to a more conservative Bayes estimate, there may be situations when Bayes and other methods underestimate the value of this rare event. This result is demonstrated through an example in developmental toxicology.

The conclusion of this paper is not necessary to recommend the minimax or any other estimator in all situations when there is a zero response. Rather, the goal is to increase awareness and recommend that more caution should be taken when any single method is used to estimate the success probability when sample shows zero occurrence. The choice of the estimate should to a large extent depend on which kind of optimality is judged to be most appropriate for the case in question.

## References

- Bailey, R. T. (1997). Estimation from zero-failure data. *Risk Analysis*, 17, 375-380.
- Basu, A. P., Gaylor, D. W. & Chen, J. J. (1996). Estimating the probability of occurrence of tumor for a rare cancer with zero occurrence in a sample. *Regulatory Toxicology and Pharmacology*, 23, 139-144.
- Bickel, P.J., & Doksum, K. A.(2001). *Mathematical Statistics*. (2nd ed.) Oakland, CA: Holden-Day.
- Calabrese, E. J., & Baldwin, L. A. (2000). Chemical hormesis: Its historical foundations as a biological hypothesis. *Human and Experimental Toxicology*, 19, 2-31.
- Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. (2nd ed.). NY: Chapman and Hall.
- Chew, V. (1971) Point estimation of the parameter of the binomial distribution. *The American Statistician*, 25, 47-50.
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.
- Good, I. J. (1965). *The estimation of probabilities: An essay on modern Bayesian methods*. Research Monograph No. 30. Cambridge: The M.I.T. Press, p. 15-19.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*, Volume 2. (2<sup>nd</sup> ed.) NY: Wiley.
- Kochhar, D. M., Shealy, Y. F., Penner, J. D., & Jiang, H. (1992). Retinamides: Hydrolytic conversion of retinoylglycine to retinoic acid in pregnant mice contributes to teratogenicity. *Teratology*, 45, 175-185.
- Louis, T. A.(1981). Confidence intervals for a binomial parameter after observing no successes. *The American*, 35, 154.
- Neuman, J. von, & Morgenstern, O. (1994). *Theory of games and economic behavior*. (3rd ed.). Princeton, NJ: Princeton University Press.
- Pearson, E. S., & Hartley, H. O. (1956). *Biometrika tables for statisticians*. London: Cambridge University Press.
- Razzaghi, M., & Loomis, P. (2001). The concept of hormesis in developmental toxicology. *Human and Ecological Risk Assessment*, 7, 933-942.
- Roussas, G. G. (1997). *A course in mathematical statistics*. (2<sup>nd</sup> ed.) NY: Academic Press.
- Typaldos, Z. A., & Brimley, D. E. (1962). Point estimation of reliability from results of a small number of trials. *Memorandum RM-3044-PR*, Santa Monica, CA: The Rand Corporation.
- SAS(1999). *Statistical analysis system*, Version 8. Cary, NC: SAS Institute.

## Null Distribution Of The Likelihood Ratio Statistic For Feed-Forward Neural Networks

Douglas Landsittel  
Dept. of Biostatistics  
University of  
Pittsburgh

Harshinder Singh  
Department of Statistics  
West Virginia University

Vincent C. Arena  
Dept. of Biostatistics  
University of Pittsburgh

Stewart J. Anderson  
Dept. of Biostatistics  
University of Pittsburgh

---

Despite recent publications exploring model complexity with modern regression methods, their dimensionality is rarely quantified in practice and the distributions of related test statistics are not well characterized. Through a simulation study, we describe the null distribution of the likelihood ratio statistic for several different feed-forward neural network models.

Key words: degrees of freedom, model complexity, chi-square distribution.

---

### Introduction

Neural networks have become a popular regression method for classification and prediction of high-dimensional and/or highly non-linear data (Ripley, 1994). Their appeal in such circumstances is due to their implicitly non-linear model structure, which does not require the user to explicitly define the presence, or degree, of interactions and non-linear terms, and subsequent ability to universally approximate any function (Ripley, 1996). In cases where complex models are needed to fit the underlying associations, but the nature of those associations is not well understood, neural networks are hypothesized to offer a more effective approach to classification. Other consequences of this implicit non-linearity, however, are 1) the propensity of neural networks to over-fit the training data, and 2) the inability to equate the number of model parameters with the effective model dimension.

Other studies have rigorously investigated the issue of model complexity, both specifically for neural networks, and more generally for non-parametric and non-linear regression models. Hastie and Tibshirani (1990), and Loader (1999) calculated degrees of freedom for scatterplot smoothers, local regression, and other nonparametric models using the trace of the hat matrix. For more complex models or model selection procedures, where the hat matrix cannot be explicitly specified, Ye (1998) proposes the generalized degrees of freedom, which estimates the hat matrix diagonal based on the sensitivity of fitted values to changes in observed response values. Hodges and Sargent (2001) extended degrees of freedom to random effects, hierarchical models, and other regression methods (and show a connection to Hastie & Tibshirani, 1990; and Ye, 1998) using a re-parameterization of the trace of the hat matrix.

More specific to neural networks, Moody (1992) and others (Ripley, 1995; Liu, 1995; Amari & Murata, 1993; Murata, Yoshizawa, & Amari, 1991) calculated the effective number of model parameters based on approximating the test set error as a function of the training set error plus model complexity. Other methods (as summarized by Ripley, 1996; and Tetko, Villa, & Livingstone, 1996) include cross-validation, and eliminating variables based on small (absolute) parameter values, or variables with a small effect on predicted values (i.e. sensitivity methods). Bayesian approaches have also been proposed (Ripley, 1995; Ripley, 1996; Paige & Butler,

---

Douglas Landsittel ([landsittel@upci.pitt.edu](mailto:landsittel@upci.pitt.edu)) is Research Assistant Professor, Biostatistics Dept., University of Pittsburgh, and Statistician, Pittsburgh Cancer Institute. Harshinder Singh ([his6@cdc.gov](mailto:his6@cdc.gov)) is Research Professor, Statistics Department, West Virginia University, and Senior Researcher, Biostatistics Branch, NIOSH/HELD. Vincent C. Arena ([arena@pitt.edu](mailto:arena@pitt.edu)) and Stewart Anderson ([andersons@nsapb.pitt.edu](mailto:andersons@nsapb.pitt.edu)) are Associate Professor, Biostatistics Department, University of Pittsburgh.

2001) for model selection with neural networks. Implementation of such methods, however, has been limited by either computational issues, dependence on the specified test set, or lack of distributional theory.

To our knowledge, no previous studies have directly investigated the distribution of the likelihood ratio statistic with neural networks. In this study, simulations are conducted to empirically describe the distribution of the likelihood ratio statistic under the null assumption of the intercept model (versus the alternative of at least one non-zero covariate parameter). All simulations are conducted with a single binary response; in contrast, the previously cited literature primarily focuses on continuous outcomes. In cases where the likelihood ratio can be adequately approximated by a chi-square distribution, the degrees of freedom can be used to quantify neural network model complexity under the null. Derivation of the test statistic null distribution is pursued through simulation approaches, rather than theoretical derivations, because of the complexity of the network response function and the lack of maximum likelihood or other globally optimal estimation.

The two main objectives of this simulation study are to 1) verify that the chi-square distribution provides an adequate approximation to the empirical test statistic distribution in a limited number of simulated cases, both for the test of independence and tests of nested models, and 2) quantify how the distribution and number of covariates, and the number of hidden units affects model degrees of freedom. Adequacy of the chi-square approximation will be judged by how close the  $\alpha$ -level based on the simulation distribution (i.e. the percent of the test statistic distribution greater than the corresponding chi-square quantile) is to various percentiles of the chi-square distribution. The variance, which should be approximately twice the mean under a chi-square distribution, is also displayed for each simulation condition.

### Methodology

#### A Feed-Forward Neural Network Model

This study is restricted to feed-forward models, which are the most common type of neural networks implemented in classification of

single dichotomous outcomes. We assume that  $y$  follows a Bernoulli distribution;  $x$ -values can follow any distribution, but are scaled to the interval  $[0,1]$  before fitting the model. Without doing so, the initial weights of the network would have to account for differences in magnitude, as would the process of weight decay (described later).

The predicted value,  $\hat{y}$ , for the  $k^{\text{th}}$  observation, with covariate values (or inputs)  $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{pk})$ , is given by

$$\hat{y}_k = f(v_0 + \sum_{j=1}^H v_j f\{w_{j0} + \sum_{i=1}^p w_{ji}x_{ik}\}) \quad (1)$$

(Ripley, 1996), where  $f(x)$  is the logistic function,  $\frac{1}{(1 + e^{-x})}$ . Each logistic function of the

weight sum of the data,  $f\{w_{j0} + \sum_{i=1}^p w_{ji}x_{ik}\}$ , is referred to as the  $j^{\text{th}}$  hidden unit. The predicted response of the neural network is calculated as a linear combination of these hidden unit values; the parameters  $v_0, v_1, \dots, v_H$  are referred to as the connections between the hidden and output layer. Each set of parameters  $w_{j1}, w_{j2}, \dots, w_{jp}$  then represents the weights of the  $p$  covariate values specific to the  $j^{\text{th}}$  hidden unit, or the connections between the input and hidden layer. One implication of this non-linear model structure is that none of the parameter values directly corresponds to any specific main effect or interaction.

Model fitting is typically accomplished through the procedure of back-propagation (Rumelhart, et al., 1995), where model parameters are iteratively updated using a gradient descent-based algorithm. We used the nnet function by Ripley in S-Plus (Venables & Ripley, 1997) to fit all neural network models in this study. The error criteria for dichotomous outcomes, namely minimization of

$$E = \sum_{k=1}^n [y_k \log \frac{y_k}{\hat{y}_k} + (1 - y_k) \log \frac{1 - y_k}{1 - \hat{y}_k}] \quad (2)$$

with respect to the parameters of interest is equivalent to finding global maxima of the corresponding likelihood function.

This study also incorporated weight decay, which is almost universally used to improve optimization and generalization. Rather than minimizing  $E$  in Equation 2, the fitting algorithm is applied to minimize

$$E + \lambda \sum_{j=1}^H \sum_{i=1}^p [v_j^2 + w_{ji}^2], \quad (3)$$

and thus penalize the network for large parameter values. To determine the magnitude of  $\lambda$  for dichotomous outcomes, Ripley (1996) recommended exploration in the range of  $[0.001, 0.1]$ , which is based on Bayesian arguments and the range of the logistic function. For this study, we utilized  $\lambda = 0.01$  for most simulations; additional simulations were also conducted with  $\lambda = 0.10$ .

#### Likelihood Ratio Test of Independence

The likelihood ratio statistic for testing model independence with neural networks corresponds to the usual expression from logistic regression,

$$D = 2 \left\{ \sum_{k=1}^n [y_k \log \hat{y}_k + (1 - y_k) \log(1 - \hat{y}_k)] - [n_1 \log n_1 + n_0 \log n_0 - n \log n] \right\}, \quad (4)$$

where  $n_1 = \sum_{k=1}^n y_k$ ,  $n_0 = n - n_1$ , and  $\hat{y}_k$  is calculated from Equation 1 (Cox & Snell, 1989). As opposed to the logistic model, however, the  $\hat{y}_k$  do not typically represent the maximum likelihood estimates, rather they represent only locally optimal parameter values. A primary aim of this study will therefore be to assess the adequacy of the chi-square distribution for approximating the null distribution of likelihood ratio test (of model independence) with neural networks.

This study will also investigate the null test statistic distribution for differences between nested models. Denoting  $D_R$  and  $D_F$  as the

likelihood ratio statistics for model independence of the reduced and full models, respectively,  $D_F - D_R$  gives the usual likelihood ratio test for significance of the covariates in the full but not the reduced model.

#### A Simulation Study

To investigate the null distribution (i.e. under the intercept model) of the likelihood ratio statistic (Equation 3), we simulated random data with the following characteristics. Covariate values  $\{x_{ik}\}$  were simulated with  $n = 2,000$  observations and between two and five covariates. Covariates and a single binary outcome were first randomly generated from a Bernoulli distribution with  $\Pr[x_{ik}=1] = 0.5$  and  $\Pr[y_k=1] = 0.5$ . The first two covariates,  $x_1$  and  $x_2$ , were simulated with 75 percent concordance, i.e.  $\Pr[x_{2k}=1 | x_{1k}=1] = 0.75$  and  $\Pr[x_{2k}=0 | x_{1k}=0] = 0.75$ ; all other Bernoulli covariates were independently generated. Covariates were then generated from a standard normal distribution with a correlation of 0.50 between  $x_{1l}$  and  $x_{l2}$ ; all other normal covariates were independently generated. All simulations included the two correlated (either Bernoulli or standard normal) variables and 0 to 3 independent covariates. Neural network models with 2, 5, and 10 hidden units were fit to the simulated data. Model fitting incorporated weight decay ( $\lambda = 0.01$  or 0.10) (as previously-described).

Means and variances of the simulated likelihood ratio statistics,  $D_s$ , are displayed for each simulation condition. Each simulated distribution (for a given number of inputs and hidden units) was then associated with the chi-square distribution having degrees of freedom equal to the mean (simulated) likelihood ratio ( $\bar{D}$ ). Simulated  $\alpha$ -levels ( $\alpha_q^{(S)}$ ) were then defined as the percentage of simulated values greater than  $q^{\text{th}}$  percentile of the corresponding chi-square distribution. For instance, the nominal  $\alpha$ -level for the simulated distribution is given by

$$\alpha_{0.05}^{(S)} = P[D \geq \chi_{0.05}^2(\bar{D})]. \quad (5)$$

Simulated  $\alpha$ -levels will then be compared to the chi-square percentiles at significance levels of 0.75, 0.50, 0.25, 0.10, and 0.05. Q-Q plots will

also be presented to quantify agreement with the appropriate chi-square distribution.

Results

Simulations were first conducted to investigate the null distribution of the likelihood ratio for testing model independence with strictly binary input variables (Table 1, following page). Results indicate reasonable agreement between the simulated  $\alpha$ -levels and the corresponding percentiles of the chi-square distribution. The average simulated  $\alpha$ -levels, across the 12 conditions, were all within 0.02 of the expected values. Individually, none of the simulated  $\alpha$ -levels varied more than 0.04 from the corresponding chi-square percentile. Based on this correspondence between the simulated results and the chi-square distribution, the mean likelihood ratio can be interpreted as model degrees of freedom.

The Q-Q plot of the likelihood ratio statistic (for testing model independence) with 5 binary inputs and 10 hidden units is displayed in Figure 1, which is generally representative of the other Q-Q plots. The diagonal line through  $x = y$  represents perfect agreement between the two distributions. The somewhat greater than expected test statistic variance (66.8 as opposed to twice the mean, which is 57.6) is evidenced by larger values of the statistic at the upper end of the distribution; slightly lower test statistic values were observed at the lower end of the distribution. This deviation in the variance, however, led to only slightly liberal  $\alpha$ -levels.

The degrees of freedom varied between approximately 3 for 2 binary inputs, to almost 30 for five binary inputs (with 10 hidden units). The number of hidden units seemed to have a greater effect on the resulting degrees of freedom with 5 inputs than with 2-4 inputs. The model with 5 inputs and 10 hidden units had nearly twice the degrees of freedom as the model with 5 inputs and 2 hidden units.

Table 2 (next page) displays simulation results for comparing the reduced model with between 2 and 4 binary covariates to the full model with all 5 binary covariates. The reduced models were specified by removing  $x_5$  to  $x_3$  in reverse order. For instance, a model reduced to 3

covariates,  $\{x_1, x_2, x_3\}$ , would be compared to the full model with all 5 covariates.

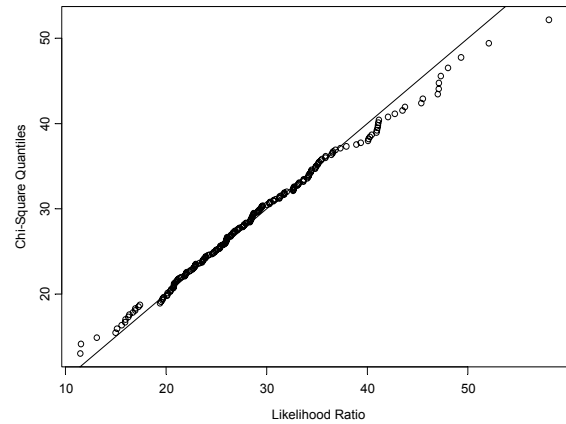


Figure 1. Q-Q Plot of the Likelihood Ratio with 5 Binary Covariates and 10 Hidden Units

The average simulated  $\alpha$ -levels, across the 12 conditions, were all within 0.02 of the expected values. With one exception (2 hidden units and 4 inputs in the reduced model), none of the simulated  $\alpha$ -levels individually varied more than 0.04 from the corresponding chi-square percentile, and most simulated results were within 0.02 of the chi-square percentile.

The degrees of freedom varied between approximately 5 when adding 1 binary input to the reduced model with 4 inputs (and 2 hidden units), to 26 when adding 3 binary inputs to the reduced model with 2 inputs (and 10 hidden units). The number of hidden units seemed to have a greater effect on the resulting degrees of freedom using the reduced model with 4 inputs. Testing the addition of a single binary input to the reduced model with 4 inputs equated to 15 degrees of freedom with 10 hidden units, as opposed to 5 degrees of freedom with 2 hidden units.

Table 3 (following page) presents simulation results for the case of standard normal covariates. Results again indicated reasonable agreement between the simulated  $\alpha$ -levels and the corresponding percentiles of the chi-square distribution. The average simulated  $\alpha$ -levels, across the 12 conditions, were all within 0.02 of the expected values. Individually, all of the simulated  $\alpha$ -levels were within approximately 0.05 of the corresponding chi-square percentile plot in Figure 1 was also generally representative of the Q-Q plots for testing nested models.



Table 1. Likelihood Ratio Statistic for Model Independence with Binary Inputs

Inputs	Hidden Units	Likelihood Ratio		Simulated $\alpha$ -levels				
		Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	2	2.8	6.2	0.715	0.535	0.245	0.090	0.055
	5	2.8	6.1	0.720	0.530	0.240	0.085	0.050
	10	2.8	6.1	0.720	0.530	0.240	0.085	0.050
3	2	5.9	13.5	0.700	0.480	0.285	0.120	0.060
	5	6.2	14.3	0.710	0.485	0.270	0.095	0.060
	10	6.3	14.3	0.710	0.480	0.270	0.100	0.060
4	2	10.5	22.6	0.730	0.495	0.265	0.105	0.040
	5	13.7	34.4	0.735	0.490	0.245	0.105	0.070
	10	13.8	34.5	0.740	0.490	0.245	0.105	0.070
5	2	15.6	33.3	0.750	0.520	0.235	0.125	0.080
	5	27.4	61.7	0.755	0.475	0.240	0.115	0.065
	10	28.8	66.8	0.740	0.490	0.265	0.125	0.065
Mean Simulated $\alpha$ -levels				0.727	0.500	0.254	0.105	0.060

Table 2. Likelihood Ratio Statistic for Nested Models with Binary Inputs

Reduced Model	Hidden Units	Likelihood Ratio		Simulated $\alpha$ -levels				
		Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 inputs	2	12.8	26.1	0.760	0.515	0.240	0.105	0.065
	5	24.6	51.7	0.755	0.490	0.240	0.105	0.070
	10	26.0	56.4	0.750	0.455	0.265	0.110	0.085
3 inputs	2	9.7	23.2	0.750	0.500	0.285	0.110	0.060
	5	21.2	43.6	0.755	0.475	0.255	0.105	0.070
	10	22.6	47.5	0.745	0.490	0.265	0.100	0.075
4 inputs	2	5.1	18.1	0.695	0.535	0.305	0.145	0.090
	5	13.7	26.3	0.750	0.490	0.240	0.095	0.055
	10	15.1	28.2	0.750	0.495	0.250	0.090	0.050
Mean Simulated $\alpha$ -levels				0.746	0.494	0.261	0.107	0.069

Table 3. Likelihood Ratio Statistic for Model Independence with Standard Normal Inputs

Inputs	Hidden Units	Likelihood Ratio		Simulated $\alpha$ -levels				
		Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	2	9.1	19.3	<b>0.750</b>	0.540	0.290	0.105	0.045
	5	21.8	50.8	0.735	0.500	0.270	0.100	0.045
	10	39.4	101.3	0.725	0.540	0.280	0.135	0.040
3	2	13.8	24.2	0.765	0.555	0.250	0.085	0.030
	5	34.9	65.6	0.760	0.505	0.270	0.095	0.040
	10	69.4	133.7	<b>0.755</b>	0.540	0.250	0.075	0.025
4	2	19.1	31.0	0.795	0.520	0.250	0.085	0.040
	5	47.5	84.7	0.775	0.525	0.255	0.075	0.045
	10	100.4	158.1	0.800	0.530	0.220	0.075	0.030
5	2	23.5	49.6	0.765	0.495	0.240	0.110	0.045
	5	61.3	110.9	0.775	0.495	0.225	0.095	0.025
	10	128.5	206.4	0.780	0.520	0.205	0.085	0.025
Mean Simulated $\alpha$ -levels				0.765	0.522	0.250	0.093	0.036

Table 4. Likelihood Ratio Statistic for Nested Models with Standard Normal Inputs

Reduced Model	Hidden Units	Likelihood Ratio		Simulated $\alpha$ -levels				
		Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 inputs	2	14.4	54.1	<b>0.705</b>	0.510	0.315	0.150	0.090
	5	39.5	150.3	0.705	0.540	0.320	0.155	0.085
	10	88.1	262.5	0.710	0.505	0.300	0.140	0.100
3 inputs	2	9.7	52.5	0.660	0.510	0.340	0.215	0.135
	5	26.4	135.8	0.685	0.515	0.340	0.210	0.145
	10	58.1	266.0	<b>0.665</b>	0.505	0.355	0.230	0.130
4 inputs	2	4.4	56.6	0.605	0.515	0.400	0.245	0.195
	5	13.8	152.8	0.615	0.535	0.350	0.260	0.205
	10	27.1	260.3	0.630	0.500	0.385	0.270	0.230
Mean Simulated $\alpha$ -levels				0.664	0.515	0.345	0.208	0.146

Table 5. Likelihood Ratio Statistic for Nested Models with Standard Normal Inputs and Weight Decay of 0.10

Reduced Model	Hidden Units	Likelihood Ratio		Simulated $\alpha$ -levels				
		Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 inputs	2	10.8	21.6	<b>0.780</b>	0.550	0.235	0.120	0.060
	5	35.2	95.1	0.710	0.495	0.255	0.160	0.090
	10	73.0	158.5	0.725	0.525	0.255	0.125	0.060
3 inputs	2	7.5	20.1	0.745	0.520	0.280	0.105	0.075
	5	24.1	94.5	0.695	0.500	0.315	0.185	0.120
	10	51.9	181.0	<b>0.695</b>	0.520	0.305	0.165	0.090
4 inputs	2	4.1	21.3	0.585	0.450	0.365	0.210	0.130
	5	12.3	72.9	0.655	0.515	0.380	0.210	0.135
	10	25.6	134.6	0.675	0.520	0.350	0.240	0.140
Mean Simulated $\alpha$ -levels				0.696	0.511	0.304	0.169	0.100

The degrees of freedom varied between approximately 9 for 2 binary inputs (with 2 hidden units), to 128 for five binary inputs (with 10 hidden units). The number of hidden units greatly affected the resulting degrees of freedom for all simulated cases. The model with 5 hidden units corresponded to approximately twice the degrees of freedom as the model with 2 hidden units, and half the degrees of freedom as the model with 10 hidden units.

The Q-Q plot of the likelihood ratio statistic (for testing model independence) with 5 standard normal inputs and 10 hidden units is displayed in Figure 2. It is generally representative of the other Q-Q plots. The somewhat lesser than expected test statistic variance (206.4 as opposed to twice the mean, which is 257.0) is evidenced by smaller values of the statistic at the upper end of the distribution. The nominal  $\alpha$ -level were subsequently somewhat conservative.

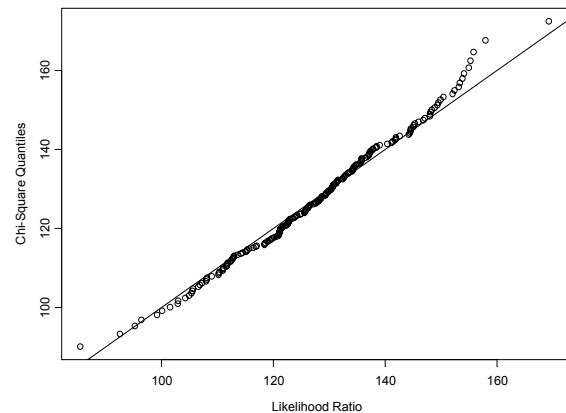


Figure 2. Q-Q Plot of the Likelihood Ratio with 5 Standard Normal Covariates and 10 Hidden Units

Table 4 (previous page) displays simulation results for comparing the reduced model with between 2 and 4 standard normal covariates to the full model with all 5 standard normal covariates. These results, as opposed to previous simulations, do not reflect correspondence to a chi-square distribution. The simulated distributions for testing nested models with continuous covariates are far more skewed; the variance was often 4 or more times greater than the mean (in contrast to the expected 1:2 mean-variance ratio). On average, across the 12 conditions, the difference between simulated  $\alpha$ -levels and chi-square percentiles was approximately 10 percent.

To address the substantial discrepancies in Table 4, simulations were rerun using a weight decay of 0.10. Results in Table 5 show a slightly better correspondence to the chi-square distribution under some conditions, but still reflect far greater variability in the test statistic, and subsequently large differences from the chi-square percentiles. The nominal 0.05  $\alpha$ -level, for instance, was between 0.06 and 0.09 for testing the reduced model with 2 standard normal covariates, but was at least 13 percent for testing the reduced model with 4 covariates.

### Conclusion

The chi-square distribution appears to provide an adequate approximation to the null distribution (assuming no association between covariates and response) for likelihood ratio tests of independence with feed-forward neural networks. Tests between nested models are approximately chi-square for strictly binary inputs, but not for standard normal covariates. Apart from significance testing, one contribution of these simulations is to quantify the model complexity (under the null) for various neural network models. Although the implicitly non-linear nature of neural networks is commonly known, specifically quantifying the effective number of model parameters remains a difficult task.

These simulations illustrate that even a neural network with only 5 strictly binary inputs (and ten hidden units) can implicitly fit nearly 29 degrees of freedom. Testing the significance of a single binary input, against the reduced model with 4 binary inputs, equates to approximately 15 degrees of freedom. Neural networks with continuous covariates resulted in even greater model complexity; the neural network with 5 standard normal covariates and 10 hidden units equated to approximately 129 degrees of freedom.

The degrees of freedom with strictly binary inputs can be conceptualized as the number of main effects and interaction terms fit by the neural network model; other non-linear functions of a binary term are still 0 or 1, and therefore not relevant. In a related technical report (Landsittel, et al., 2002a), we explored these same models (of strictly binary data) using globally optimal parameter estimates; numerous initial weights were implemented to conduct a grid search of the

likelihood surface. In that study, the degrees of freedom was equal to the number of covariate patterns minus one for the intercept (i.e.  $2^p-1$ , where  $p$  is the number of parameters) given a sufficient number of hidden units. For simulations where there was an insufficient number of model parameters to fit the saturated model (i.e. the number of parameters was less than  $2^p-1$ ), the degrees of freedom was greater than the number of model parameters, but less than the number of covariate patterns. In the current study, based on the usual algorithm which picks only one randomly chosen set of initial parameters, the degrees of freedom was always less than the number of covariate patterns. For instance, 2 binary inputs equates to 2 main effects and 1 interaction term yielding 3 degrees of freedom. The simulated degrees of freedom subsequently equaled 3.0 in the previously-described technical report (based on globally optimal models), and was slightly less, at 2.8, in this current study.

The neural network models with standard normal covariates implicitly fit not only main effects and interactions, but also an indeterminate number of non-linear terms (of an indeterminate nature). This is evidenced by the greater degrees of freedom associated with standard normal covariates (i.e. Table 3 versus Table 1). Consider, for instance, the Taylor series expansion (using the first  $q$  terms) of the neural network response function for the  $k^{\text{th}}$  observation with a single continuous covariate.

$$\begin{aligned} \log \text{it}(\hat{y}_k) = & v_0 + \sum_{j=1}^H v_j f(w_{j0}) + x_k \left( \sum_{j=1}^H v_j w_{j1} f'(w_{j0}) \right) \\ & + \frac{1}{2} x_k^2 \left( \sum_{j=1}^H v_j w_{j1}^2 f''(w_{j0}) \right) \\ & + \dots + \frac{1}{q!} x_k^q \left( \sum_{j=1}^H v_j w_{j1}^q f^{(q)}(w_{j0}) \right) \end{aligned} \tag{6}$$

No clear correspondence can be derived between the number of parameters and the number of implicitly fit non-linear terms. This approximation underscores both the implicitly nonlinear structure and the lack of interpretable coefficients. Each expansion term is a function of multiple network parameters and, with the exception of  $v_0$  (the hidden layer intercept term), each network

parameter is involved in calculating multiple expansion terms.

The results of this simulation reflect the unpredictable nature of model complexity with neural networks. The degrees of freedom varies both according to the number of input variables and the distribution of these covariates, as well as the number of hidden units. Furthermore, the degrees of freedom will also depend significantly on other issues not investigated here, such as the underlying association (all simulations here were under the null), use of additional training modifications (e.g. model averaging or early stopping of training based on a test set), and further variations in the covariate distributions. This would imply that, from these simulations, we can still only specify the appropriate degrees of freedom in very limited cases.

To address this limitation, we are currently investigating an explicit approach to calculate degrees of freedom with neural networks and dichotomous outcomes. The approach is based on a simple modification to Ye's (1998) procedure for generalized degrees of freedom in the continuous case. The resulting measure for a binary outcome corresponds to Fay's range of influence (ROI) statistic for logistic regression. In a recent commentary (Landsittel, et al., 2002), we empirically show that Fay's ROI statistic asymptotically corresponds to the hat matrix diagonal, and therefore (the sum of these ROI statistics) provides a potential measure of degrees of freedom. Additional simulations will focus on connecting this statistic to the mean likelihood ratio over simulated distributions with neural networks.

In addition to the methods employed here, numerous other training modifications, such as committees of networks, or early stopping of training based a test set, are frequently used and do affect model complexity. Additional simulations (not shown here) indicated that neither network committees nor early stopping lead to correspondence with a chi-square distribution. Greater values of weight decay, or other modifications to model fitting, may lead to a better correspondence with chi-square percentiles in the case of testing nested models with standard normal covariates. In addition to slight improvement of the chi-square approximation, increasing the weight decay tends to reduce the mean likelihood

ratio implicitly fit under the null. Further variations on neural network models, such as other covariate distributions, will likely effect the model complexity in an unpredictable manner. These issues can be better explored once an explicit measure is derived for calculating degrees of freedom with a binary outcome.

Although other methods exist for inference and quantifying model complexity with neural networks, these approaches are not widely implemented because of associated computational issues (see Introduction). Use of the likelihood ratio statistic provides a more widely utilized approach, which is easily calculated from the observed and predicted response values (using common statistical programs such as S-Plus). Results of this approach can also be easily interpreted by applied researchers.

#### References

- Amari, S. and Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, 5, 140-153.
- Cox, D.R. and Snell, E.J. (1989). *Analysis of binary data*. New York, NY: Chapman and Hall, 26-102.
- Faraggi, D., & Simon, R. (1995). Maximum likelihood neural network prediction models. *Biometrical Journal*, 37(6), 713-725.
- Fay, M. (2002). Measuring a binary response's range of influence in logistic regression. *The American Statistician*, 56(1), 5-9.
- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. New York: Chapman and Hall, 150-152.
- Hodges, J.S., & Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2), 367-379.
- Landsittel, D., Singh, H., Arena V.C. (2002a). Likelihood ratio test of independence for binary data with neural networks. *Technical Report Series – Methods #37*, Department of Biostatistics, University of Pittsburgh.
- Landsittel, D., Singh, H., Arena, V.C., & Anderson, S. (2002). Comment on "Measuring a binary response's range of influence in logistic regression." *The American Statistician*.

Liu, Y. (1995). Unbiased estimate of generalization error and model selection in neural networks. *Neural Networks*, 8(2), 215-219.

Moody, J.E. (1992). The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.) *Advances in neural information processing systems 4*. San Mateo, CA: Morgan Kaufmann, 847-854.

Murata, N., Yoshizawa, S., & Amari, S. (1991). A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen, K. Makisara, O. Simula, & J. Kangas (Eds.) *Artificial neural networks*. North Holland: Elsevier Science Publishers, 9-14.

Paige, R. L. and Butler, R. W. (2001). Bayesian inference in neural networks. *Biometrika*, 88(3), 623-641.

Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press, 143-180.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D.E. Rumelhart (Eds.) *Backpropagation: Theory, architectures, and applications*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1-34.

Tetko, I.V., Villa, A.E., & Livingstone, D.J. (1996). Neural network studies 2: Variable selection. *Journal of Chemical Informatics and Computer Science*, 36(4), 794-803.

Venables, W. N., & Ripley, B. D. (1997). *Modern applied statistics with S-Plus*. New York: Springer-Verlag, 337-341.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441), 120-131.

## A Simulation Study Of The Impact Of Forecast Recovery For Control Charts Applied To ARMA Processes

John N. Dyer  
Dept. Of Information Systems  
& Logistics  
Georgia Southern University

B. Michael Adams  
Dept. Of Information Systems,  
Management Science, & Statistics  
University Of Alabama

Michael D. Conerly  
Dept. Of Information Systems,  
Management Science, & Statistics  
University Of Alabama

---

Forecast-based schemes are often used to monitor autocorrelated processes, but the resulting forecast recovery has a significant effect on the performance of control charts. This article describes forecast recovery for autocorrelated processes, and the resulting simulation study is used to explain the performance of control charts applied to forecast errors.

Key words: Autocorrelation, Forecast Recovery, Box-Jenkins, Quality, Simulation, Statistical Process Control, Statistics.

---

### Introduction

Many traditional control charts were developed under the assumption that the measurements resulting from the in-control process are independent and identically distributed (iid) random variables.

Recently, many advances in measurement technology and sampling frequency yield sample measures that are not independently distributed. Hence, an alternative to the traditional control charting approach is to utilize a forecast-based monitoring scheme, which involves identifying the proper time-series model characterizing the process, obtaining the appropriate Box-Jenkins one-step-ahead forecast of process observations,

and then applying traditional control charts to forecast errors (Alwan & Roberts, 1988; Wardell, Moskowitz, & Plante, 1994; Lin & Adams, 1996; Lu & Reynolds, 1999a; Lu & Reynolds, 1999b; Lu & Reynolds, 2001). If the assumed time-series model is correct, the forecast errors are iid normal random variables. Hence, the errors perform in a manner predictable through traditional control charting techniques, enabling monitoring for detection of step-shifts in the process mean level.

One problematic characteristic of forecast-based monitoring schemes is the phenomenon of forecast recovery; that is, the process forecasts recover quickly from process disturbances. Hence, the resulting forecast errors also recover quickly. This article describes models for autocorrelated data and the impact of forecast recovery for three special cases of the general autoregressive moving average (ARMA) model, and investigates the impact of forecast recovery on the Individuals, Exponentially Weighted Moving Average (EWMA), and the Combined EWMA-Shewhart (CES) control charts applied to forecast errors resulting from the ARMA models. A description of the simulation study is also provided. Recommendations are provided that will enable the practitioner to more readily identify the most appropriate control chart for use in monitoring various ARMA processes.

---

John N. Dyer is an Assistant Professor of Decision Sciences. His area of specialty is statistical process control with an emphasis in autocorrelated processes and forecast-based control charting. He is a member of ASA, ASQ, and DSI. Email: [jdyer@gasou.edu](mailto:jdyer@gasou.edu), Phone: 912.681.5223, Fax: 912.681.0710. B. Michael Adams is an Associate Professor of Statistics. His primary areas of research include quality control and statistical process control. He is a member of ASA and ASQ. Email: [badams@cba.ua.edu](mailto:badams@cba.ua.edu) Michael D. Conerly is a Professor of Statistics. His areas of research include quality control and statistical process control. He is a member of ASA and ASQ. Email: [mconerly@cba.ua.edu](mailto:mconerly@cba.ua.edu).

### Methodology

When control chart performance has been evaluated, the average run length (ARL) has

typically been used to quantify performance of the chart. The ARL is defined as the average number of time periods until the control chart signals. When the process is in-control, this is the expected time until a false-alarm. When the process shifts out-of-control, the ARL measures the expected time to detect the shift. The desired chart is one that simultaneously provides large in-control ARLs and low out-of-control ARLs. An alternative performance criterion is the cumulative distribution function (CDF). The CDF measures the cumulative proportion or percent of signals given by the  $i^{\text{th}}$  period following the shift. It should be noted that the CDF completely characterizes the run length distribution, while the ARL is only the mean. Additionally, the median run length (MRL) can be used in conjunction with the ARL and CDF since it is a better measure of central tendency for skewed distributions such as the run length distribution. The MRL is defined as the median ( $50^{\text{th}}$  percentile) number of time periods until the control chart signals. The desired chart is one with a high probability of early detection of a shift. In most cases, a trade-off between obtaining a low out-of-control ARL and high probability of early detection results.

The impact of forecast error recovery on ARLs has been discussed (Adams, Woodall, & Superville; 1994; Superville & Adams, 1994), and the CDF technique has been recommended as a meaningful criterion for evaluating the performance of charts on forecast errors. In light of forecast recovery, both ARL and CDF performance for step-shifts in the process mean were evaluated (Lin & Adams, 1996) on the Individuals chart, the exponentially weighted moving average (EWMA) chart, and the combined EWMA-Shewhart (CES), in regard to monitoring forecast errors arising from particular forecast-based monitoring schemes. It was found that the Individuals chart provides relatively high ARLs and CDFs, the EWMA provides low ARLs and CDFs, and the CES borrows the best properties from both charts, low ARLs and high CDFs. High (low) CDFs are defined as those exhibiting a high (low) probability of initial shift detection relative to competing control charts.

In this article, control chart performance results are based primarily on ARL and CDF measures, but the MRL is also provided for each chart. Standard error of the run length (SRL)

measures were provided to summarize the variability of each chart's run length distribution, as well as to give the reader an idea of the accuracy of each ARL measure. Performance results of the traditional control charts applied to forecast errors resulting from various ARMA(1,1), AR(1), and MA(1) processes with a step shift of  $c = 1\sigma_\varepsilon$  are given in Table 5.

Simulations of the performance of the Individuals, EWMA, CES control applied to the forecast errors arising from various ARMA(1,1), AR(1), and MA(1) processes in this article give some insight into the impact of forecast recovery on these traditional control charts. This insight will better enable the practitioner to choose the appropriate control chart for various ARMA processes. The control charts were designed to provide in-control ARLs of 300. The EWMA and CES control charts were designed to detect a shift of the magnitude of the sustained expected forecast error for each model. A thorough discussion of sustained forecast recovery and sustained expected forecast error is provided in the following subsections.

#### Simulation Description

The simulation programs were designed, compiled, and run in Microsoft FORTRAN PowerStation for Windows, Version 4.0, utilizing FORTRAN 90. The program for finding ARLs were also used to estimate the appropriate control limits through trial and error. The simulations conducted are as follows.

1. A series of 4,100 ARMA(2,1) variates were generated by FORTRAN MSIMSL subroutine RNARM. These variates were the simulated observations,  $Y_i$ 's, for each of the models investigated.
2. The first 100 observations were used to allow a burn-in period.
3. A step shift was induced in the simulated observations. The magnitudes of shift range from 0 to  $3\sigma_\varepsilon$  in increments of  $1\sigma_\varepsilon$ .
4. The appropriate Box-Jenkins OSA forecast and OSA forecast errors were calculated.
5. The programmed control chart monitored the forecast errors. The run lengths for the specified shift size were recorded.
6. Steps 1 through 5 were repeated 10,000 times for each model and process shift. The run length for the control chart was recorded for



each simulation repetition and the ARL was obtained based on 10,000 repetitions. For the CDF programs, the percentages of runs producing a signal within the first 300 observations following the shift were obtained.

One issue concerning the simulation should be addressed. Each program can be run to simulate a process in a zero state or steady state. Zero state provides for simulating a process from start-up, while steady state provides for simulating a process that has been running in an in-control state for some time. When simulating for control limits and Null case ARL, MRL, and CDF performance, the programs were run from zero state. When simulating the ARL, MRL, and CDF performance for a process that has experienced a shift, the programs were run from steady state.

Models for Autocorrelated Data

Two ARMA(p, q) models have been found to have application in statistical process control. The first model of interest is the ARMA(1,1). Wardell, Moskowitz, and Plante (1992) address the ARMA(1,1) model, as it is a reasonable fit to data for some manufacturing processes. The second model of interest is the ARMA(1,0), also known as the AR(1). Montgomery and Mastrangelo (1991) and Alwan and Roberts (1988) have addressed the importance of the AR(1) model in manufacturing processes. Atienga, Tang and Ang (1998) discussed a time series approach to detecting level shifts in AR(1) processes. Lastly, the ARMA(0, 1), also known as the MA(1), is considered for the sake of completion of all possible first order ARMA(p, q) models. The next section briefly discusses process shifts associated with the various time-series models before description of the models.

ARMA(1,1), AR(1), MA(1) Models & Process Shifts

In building an empirical model of an actual time-series process, the inclusion of both autoregressive and moving average terms sometimes leads to a more parsimonious model than could be achieved with either the pure autoregressive or pure moving average alone. This results in the mixed autoregressive-moving average. When both terms are mixed in first order, the resulting model is the ARMA(1, 1). The model

for an in-control ARMA(1, 1), AR(1), and MA(1) processes are given by Eq.s (1), (2), and (3) respectively,

$$Y_t = \xi + \phi Y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1} \tag{1}$$

$$Y_t = \xi + \phi Y_{t-1} + \varepsilon_t \tag{2}$$

$$Y_t = \xi - \theta \varepsilon_{t-1} + \varepsilon_t \tag{3}$$

where  $\xi$  is a constant and the sequence of  $\varepsilon_t$  ( $t = 1, 2, \dots$ ) values are independent  $N(0, \sigma_\varepsilon^2)$  random variables. The ARMA(1, 1) process is stationary for  $|\phi| < 1$  and  $|\theta| < 1$ , the AR(1) process is stationary for  $|\phi| < 1$ , and the MA(1) process is stationary for all values of  $\theta$ .

Now, suppose a step shift of size  $c$  occurs in any of the ARMA(1,1), AR(1), or MA(1) processes between time periods  $r-1$  and  $r$ , that is, the process mean suddenly changes from  $\xi$  to  $\xi+c$  at observation  $r$ . The Box-Jenkins one-step-ahead (OSA) forecasts are defined by  $\hat{Y}_t = \xi + \phi Y_{t-1} - \theta e_{t-1}$  for the ARMA(1,1) process,  $\hat{Y}_t = \xi + \phi Y_{t-1}$  for the AR(1) process, and  $\hat{Y}_t = \xi - \theta e_{t-1}$  for the MA(1) process.

The OSA forecast errors are calculated as  $e_t = Y_t - \hat{Y}_t$ , for all processes. The expected OSA forecast errors for an ARMA(1,1) process can be described mathematically as

$$E(e_t) = \begin{cases} 0 & t = 1, 2, \dots, r-1 \\ c & t = r \\ \left[ 1 - \frac{(\phi - \theta)(1 - \theta^k)}{1 - \theta} \right] c & t = r + k, k = 1, 2, \dots \end{cases} \tag{4}$$

Similar results for the AR(1) and MA(1) processes can be obtained by setting  $\theta = 0$  or  $\phi = 0$ , respectively in Eq. (4). This general representation is consistent with the special cases of the ARMA(1,1) model presented in Atienga, Tang and Ang (1998), Lin and Adams (1996), and Wardell, Moskowitz and Plante (1994).

Tables 1, 2, and 3 portray a realization of the expectation of forecast errors at time periods  $t < r$ ,  $t = r$ , and  $t > r$ , for a  $c = 1\sigma_\varepsilon$  step shift in ARMA(1,1), AR(1), and MA(1) models. These choices of models were designed by Wardell, Moskowitz, and Plante (1994) to systematically cover the region over which the ARMA series is

stationary. Although the models possessing positive autocorrelation are most likely to be encountered in manufacturing processes, those possessing negative correlation may be more prevalent in nonmanufacturing applications.

When an ARMA(p,q) process undergoes a step shift in the mean, the expected value of the forecast of the process varies for a time and then converges to a new equilibrium level (Wardell *et al.* (1994)), referred to in this paper as the sustained level of the shift. The response of the forecasts also causes the forecast errors to respond dynamically, as can be seen in Tables 1, 2, and 3. For the ARMA(1,1) model, the forecast errors react much differently, depending on the degree and direction of the first order autocorrelation,  $\rho_1$ , as well as the values of  $\phi_1$  and  $\theta_1$ . For all ARMA(p,q) models, the expected forecast error at time  $t = r$  is equal to  $c$ , but the dynamic response of the errors can vary dramatically for times  $t > r$ .

Table 1: Forecast Error Expectation for Positively Autocorrelated ARMA(1,1) Processes with a Shift of  $c = I\sigma_\epsilon$  at Time Period  $t = r$ .

Model	1	2	3	4	5	6	7	8
$\phi_1$	.950	.950	.950	.950	.475	.475	.475	-.475
$\theta_1$	.900	.450	-.45	-.90	.450	-.45	-.90	-.900
$\rho_1$	.072	.824	.971	.975	.025	.689	.737	.255
t	Expected Forecast Errors, E(e <sub>t</sub> )							
< r	.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
R+1	.95	0.50	-.40	-.85	0.98	0.08	-.38	0.58
R+2	.91	0.28	0.23	0.82	0.96	0.49	0.86	0.96
R+3	.86	0.17	-.05	-.68	0.96	0.30	-.25	0.61
R+4	.83	0.13	0.07	0.67	0.96	0.39	0.75	0.92
R+5	.80	0.11	0.02	-.55	0.96	0.35	-.15	0.64
r+44	.50	0.09	0.03	0.04	0.95	0.36	0.28	0.78
r+45	.50	0.09	0.03	0.02	0.95	0.36	0.27	0.77

For positively autocorrelated ARMA(1,1) processes, the following is observed in Table 1: The  $E(e_t)$  recovers to a value less than  $c$  for all times  $t > r$ . The recovery rate depends not only upon the values of  $\phi_1$  and  $\theta_1$ , but also upon the particular time  $t$  after the shift. Defining  $E(e_t^*)$  to be the expected sustained level of the original shift of size  $c$  resulting from an ARMA(1,1) process, Eq. (5) can be derived from Eq. (4) when  $t > r$ , as  $k \rightarrow \infty$ , and it can be shown that

Table 2: Forecast Error Expectation for Negatively Autocorrelated ARMA(1,1) Processes with a Shift of  $c = I\sigma_\epsilon$  at Time Period  $t = r$ .

$\phi_1$	.475	-.475	-	-	-.950	-.95	-.95	-.95
$\theta_1$	.900	.900	.450	-.45	.900	.450	-.45	-.90
$\rho_1$	-.255	-.737	-	-	-.975	-	-	-
			.689	.025		.971	.824	.072
t	Expected Forecast Errors, E(e <sub>t</sub> )							
< r	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00
r	1.00	1.0	1.00	1.00	1.00	1.00	1.00	1.00
r+1	1.43	2.38	1.93	1.03	2.85	2.40	1.50	1.05
r+2	1.81	3.61	2.34	1.01	4.52	3.03	1.28	1.01
r+3	2.15	4.73	2.53	1.02	6.01	3.31	1.38	1.05
r+4	2.46	5.73	2.61	1.02	7.36	3.44	1.33	1.01
r+5	2.74	6.63	2.65	1.02	8.58	3.50	1.35	1.04
r+44	5.21	14.62	2.68	1.02	19.32	3.55	1.34	1.03
r+45	5.21	14.63	2.68	1.02	19.34	3.55	1.34	1.03

Table 3: Forecast Error Expectation for AR (1) and MA(1) Processes with a Shift of  $c = I\sigma_\epsilon$  at Time Period  $t = r$ .

Model	9	10	11	12
$\phi_1$	.950	.475	-.475	-.950
$\theta_1$	.000	.000	.000	.000
$\rho_1$	.950	.475	-.475	-.950
			.000	.000
			.374	.497
				-.497
				.374
t	Expected Forecast Errors, E(e <sub>t</sub> )			
< r	0.00	0.00	0.00	0.00
r	1.00	1.00	1.00	1.00
r+1	0.05	0.53	1.48	1.95
r+2	0.05	0.53	1.48	1.95
r+3	0.05	0.53	1.48	1.95
r+4	0.05	0.53	1.48	1.95
r+5	0.05	0.53	1.48	1.95
r+44	0.05	0.53	1.48	1.95
r+45	0.05	0.53	1.48	1.95

$$E(e_t) \rightarrow \left[ 1 - \frac{(\phi_1 - \theta_1)}{(1 - \theta_1)} \right] c = E(e_t^*) \quad (5)$$

Table 4 contains values of  $E(e_t^*)$  for various combinations of  $\phi_1$  and  $\theta_1$ , hence providing a realization of the dynamic response of the forecast errors. Again, the degree of autocorrelation as well as the values of  $\phi_1$  and  $\theta_1$  determines the rate of convergence. It is obvious from Eq. (4) that  $k$  enters into the determination of  $E(e_t)$  only through  $\theta_1$ ; hence, only ARMA(1,1) and MA(1) models with nonzero  $\theta_1$  converge to  $E(e_t^*)$ . In general, it

appears that ARMA(1,1) models with large  $|\theta_1|$  converge more slowly than when  $|\theta_1|$  is small.

Models with large  $\rho_1$  tend to converge to a value of  $E(e_t^*)$  close to zero, while models with small  $\rho_1$  (i.e., close to zero) tend to quickly attain a value of  $E(e_t^*)$  close to  $c$ . For some combinations of  $\phi_1$  and  $\theta_1$ , most noticeably  $\phi_1$  positive while  $\theta_1$  negative, the  $E(e_t)$  oscillates between values less than  $c$ , until finally converging to  $E(e_t^*)$ . Again, depending upon the magnitude of  $\phi_1$  and  $\theta_1$ , the oscillation may go between positive and negative values less than  $c$  ( $\phi_1 = 0.95$ ,  $\theta_1 = -0.45$ ), or between strictly positive values less than  $c$  ( $\phi_1 = 0.475$ ,  $\theta_1 = -0.45$ ).

For negatively autocorrelated ARMA(1,1) processes, Table 2 reveals that the  $E(e_t)$  exceeds  $c$  for all times  $t > r$ . The magnitude of  $E(e_t)$  again depends on the values of  $\phi_1$  and  $\theta_1$ , as well as the time  $t$  following the shift. In most instances where  $\rho_1$  approaches zero,  $E(e_t^*)$  assumes a value trivially larger than  $c$ . In instances where  $\rho_1$  approaches negative one,  $E(e_t^*)$  often assumes a value much larger than  $c$ . Again, some oscillation among the values of  $E(e_t)$  occurs at times  $t > r$ , but not to the degree as when  $\rho_1$  is positive. Only ARMA(1,1) processes exhibiting forecast recovery, that is, positively autocorrelated processes, are further considered in this article. The ARMA(1,1) processes in Table 1 to be further considered are labeled Models 1 through 8.

For positively autocorrelated AR(1) processes, the following is observed in Table 3:  $E(e_t)$  recovers to a constant value less than  $c$  for all  $t > r$ , for all  $\rho_1$  between zero and one. Larger values of  $\phi_1$  lead to greater degrees of forecast recovery. Defining  $E(e_t^*)$  to be the expected sustained level of the original shift of size  $c$  resulting from an AR(1) process, Eq. (6) can be derived from Eq. (4) for all periods  $t > r$ , and it can be shown that

$$E(e_t^*) = (1 - \phi_1)c. \quad (6)$$

For negatively autocorrelated AR(1) processes, the following is observed in Table 3:  $E(e_t)$  increases to a constant value greater than  $c$  for all  $t > r$ , for all  $\rho_1$  between zero and negative one. Values of  $\phi_1$  closer to negative one lead to greater increases in values of the expected forecast errors. Only AR(1)

processes exhibiting forecast recovery, that is, positively autocorrelated processes, are considered in this article. The AR(1) processes to be further considered in Table 3 are labeled Models 9 and 10.

For positively autocorrelated MA(1) processes, the following is observed in Table 3:  $E(e_t)$  recovers to a value less than  $c$  for all times  $t > r$ . The recovery rate depends not only upon the value of  $\theta_1$ , but upon the particular time  $t$  after the shift. Defining  $E(e_t^*)$  to be the expected sustained level of the original shift of size  $c$  resulting from an MA(1) process, Eq. (7) can be derived from Eq. (4) when  $t > r$ , as  $k \rightarrow \infty$ , and it can be shown that

$$E(e_t) \rightarrow \left[ \frac{1}{(1 - \theta_1)} \right] c = E(e_t^*). \quad (7)$$

The degree of autocorrelation as well as the value of  $\theta_1$  determines the rate of convergence. As in the case with the ARMA(1,1),  $E(e_t)$  oscillates, converging to the value  $E(e_t^*)$ , which is less than  $c$ , for all  $t > r$ . At no time does  $E(e_t)$  exceed the value  $c$ . For negatively autocorrelated MA(1) processes, the following holds: the  $E(e_t)$  exceeds  $c$  for all times  $t > r$ . The magnitude of  $E(e_t)$  again depends on the value  $\theta_1$ , as well as the time  $t$  following the shift. The response of  $E(e_t)$  and the sustained level of the shift,  $E(e_t^*)$ , is much like that for the ARMA(1,1) model in regards to various degrees of autocorrelation. Only MA(1) processes exhibiting forecast recovery, that is, positively autocorrelated processes, are considered in this article. The MA(1) processes to be considered in Table 3 are labeled Models 11 and 12.

Table 4 contains the sustained expected forecast error values,  $E(e_t^*)$ , for various combinations of  $\phi_1$  (left most column) and  $\theta_1$  (top most row) for ARMA(1,1), AR(1), and MA(1) models, given a  $c = I\sigma_\varepsilon$  shift in the process mean level. The values  $\phi_1$  and  $\theta_1$  corresponding to the upper diagonal of Table 4 produce values of  $E(e_t)$  whose sustained level of shift is less than  $c$ . In this case, the forecast errors are said to recover. The lower diagonal region contains values of  $E(e_t)$ , whose sustained level of shift is greater than or equal to  $c$ . All entries represent combinations of  $\phi_1$  and  $\theta_1$  that result in stationary ARMA(1,1) processes.

Now consider the following example for understanding Table 4. Given an ARMA(1,1) model ( $\phi_1 = -0.15, \theta_1 = -0.65$ ) with time  $t < r$  (in-control)  $E(e_t)$  of zero, the values of  $E(e_t)$  at times  $t < r, t = r,$  and  $t > r,$  are as follows for a  $c = 1\sigma_\epsilon$  shift in the process mean:

$$E(e_t) = \begin{cases} 0 & t = 1, 2, \dots, r-1 \\ 1.00 & t = r \\ \left[ 1 - \frac{(-0.15 + 0.65)}{(1 + 0.65)} \right] = 0.70 & t = r+k \quad k \rightarrow \infty. \end{cases} \quad (8)$$

Notice that  $E(e_t)$  in Eq. (8) at time  $t > r$  is equal to 0.70 for  $k \rightarrow \infty$ . The intersection of  $\phi_1 = -0.15$  (left most column) and  $\theta_1 = -0.65$  (top most row) in Table 4 also yields the expected forecast error value of 0.70. Additionally, the relationship between the lag one autocorrelation,  $\rho_1$ , for the various combinations of  $\phi_1$  and  $\theta_1$ , and the sustained expected forecast errors,  $E(e_t^*)$ , in Table 4, is very strong and linear, with a correlation of  $r = -0.997$ . Eq. (9) provides an estimate of the sustained expected forecast errors as a function of  $\rho_1$  for most ARMA(1,1), AR(1), or MA(1) models. Again, it is obvious from this relationship that large first order autocorrelation provides for more extreme forecast recovery.

$$\hat{E}(e_t^*) = 1.04 - 1.02 \rho_1. \quad (9)$$

Considering the AR parameter alone, forecast recovery occurs for all values of  $\phi_1 > 0$ , while the most extreme sustained forecast recovery occurs for values of  $\phi_1 > 0.50$ . Considering the MA parameter alone, the sustained level of forecast error recovery never falls below  $E(e_t^*) = 0.50$ , so no value of  $\theta_1$  alone results in extreme forecast recovery. Considering both parameters, the most extreme sustained forecast recovery occurs when  $\phi_1 > 0$  while  $\theta_1 < 0$ , and in most cases in which  $\phi_1$  is large, that is,  $\phi_1 > 0.50$ , regardless of the value of  $\theta_1$ .

Results

Recall that the degree and rate of forecast recovery, as well as the time until sustained level

of forecast recovery occurs provide a source of conflict when choosing among control charts for monitoring forecast errors. Traditionally, if the ARL is used for the basis of comparison, the EWMA control chart most often provides smaller out-of-control ARLs than any other chart for small shifts, particularly when compared to the Individuals chart. However, the Individuals chart generally provides the greatest probability of obtaining a signal within the first few observations following the shift although a much larger ARL is provided. One can best understand the impact of forecast recovery by first examining chart performance applied to the AR(1) processes.

Control Charts Applied to AR(1) Models

Recall that when a shift occurs in any AR(1) process, the first forecast error following the shift appreciates the full impact of the shift,  $c$ . The forecast errors suddenly recover for all subsequent periods to a sustained level less than the original shift,  $(1-\phi_1)c$ . In contrast, the ARMA(1,1) and MA(1) processes recover gradually over time until finally converging to the sustained level less than the original shift. Depending on the particular process, oscillation may occur between values of sequential forecast errors. Since the forecast errors arising from the AR(1) process recover instantly to the sustained level of the shift, the worst performance of most control charts applied to a general ARMA(p,q) process should usually be obtained in the case of the AR(1) process for a given shift and sustained level of the shift. Performance results of the traditional control charts applied to forecast errors resulting from various ARMA(1,1), AR(1), and MA(1) processes with a step shift of  $c = 1\sigma_\epsilon$  are given in Table 5.

Table 5, Models 9 and 10, show that the EWMA control chart maintains good ARL performance relative to the Individuals chart over a wide range of AR(1) parameter values and shift sizes, but the Individuals control chart consistently provides higher probabilities of initial shift detection, particularly for larger shifts (not shown). As found by Lin and Adams (1996), the CES control chart provides out-of-control ARLs similar to those of the EWMA chart while simultaneously maintaining the high probability of an early signal provided by the Individuals chart.

	-.95	-.85	-.75	-.65	-.55	-.45	-.35	-.25	-.15	-.05	.00	.05	.15	.25	.35	.45	.55	.65	.75	.85	.95
.95	.03	.03	.03	.03	.03	.03	.04	.04	.04	.05	.05	.05	.06	.07	.08	.09	.11	.14	.20	.33	1.0
.85	.08	.08	.09	.09	.10	.10	.11	.12	.13	.14	.15	.16	.18	.20	.23	.27	.33	.43	.60	1.0	3.0
.75	.13	.14	.14	.15	.16	.17	.19	.20	.22	.24	.25	.26	.29	.33	.38	.45	.56	.71	1.0	1.7	5.0
.65	.18	.19	.20	.21	.23	.24	.26	.28	.30	.33	.35	.37	.41	.47	.54	.64	.78	1.0	1.4	2.3	7.0
.55	.23	.24	.26	.27	.29	.31	.33	.36	.39	.43	.45	.47	.53	.60	.69	.82	1.0	1.3	1.8	3.0	9.0
.45	.28	.30	.31	.33	.35	.38	.41	.44	.48	.52	.55	.58	.65	.73	.85	1.0	1.2	1.6	2.2	3.7	11
.35	.33	.35	.37	.39	.42	.45	.48	.52	.57	.62	.65	.68	.76	.87	1.0	1.2	1.4	1.9	2.6	4.3	13
.25	.38	.41	.43	.45	.48	.52	.56	.60	.65	.71	.75	.79	.88	1.0	1.2	1.4	1.7	2.1	3.0	5.0	15
.15	.44	.46	.49	.52	.55	.59	.63	.68	.74	.81	.85	.89	1.0	1.1	1.3	1.5	1.9	2.4	3.4	5.7	17
.05	.49	.51	.54	.58	.61	.66	.70	.76	.83	.90	.95	1.0	1.1	1.3	1.5	1.7	2.1	2.7	3.8	6.3	19
.00	.51	.54	.57	.61	.65	.69	.74	.80	.87	.95	1.0	1.1	1.2	1.3	1.5	1.8	2.2	2.9	4.0	6.7	20
-.05	.54	.57	.60	.64	.68	.72	.78	.84	.91	1.0	1.1	1.1	1.2	1.4	1.6	1.9	2.3	3.0	4.2	7.0	21
-.15	.59	.62	.66	.70	.74	.79	.85	.92	1.0	1.1	1.2	1.2	1.4	1.5	1.8	2.1	2.6	3.3	4.6	7.7	23
-.25	.64	.68	.71	.76	.81	.86	.93	1.0	1.1	1.2	1.3	1.3	1.5	1.7	1.9	2.3	2.8	3.6	5.0	8.3	25
-.35	.69	.73	.77	.82	.87	.93	1.0	1.1	1.2	1.3	1.4	1.4	1.6	1.8	2.1	2.5	3.0	3.9	5.4	9.0	27
-.45	.74	.78	.83	.88	.94	1.0	1.1	1.2	1.3	1.4	1.5	1.5	1.7	1.9	2.2	2.6	3.2	4.1	5.8	9.7	29
-.55	.79	.84	.89	.94	1.0	1.1	1.1	1.2	1.3	1.5	1.6	1.6	1.8	2.1	2.4	2.8	3.4	4.4	6.2	10	31
-.65	.85	.89	.94	1.0	1.1	1.1	1.2	1.3	1.4	1.6	1.7	1.7	1.9	2.2	2.5	3.0	3.7	4.7	6.6	11	33
-.75	.90	.95	1.0	1.1	1.1	1.2	1.3	1.4	1.5	1.7	1.8	1.8	2.1	2.3	2.7	3.2	3.9	5.0	7.0	12	35
-.85	.95	1.0	1.1	1.1	1.2	1.3	1.4	1.5	1.6	1.8	1.9	1.9	2.2	2.5	2.8	3.4	4.1	5.3	7.4	12	37
-.95	1.0	1.1	1.1	1.2	1.3	1.3	1.4	1.6	1.7	1.9	2.0	2.1	2.3	2.6	3.0	3.5	4.3	5.6	7.8	13	39

Table 4: Sustained Expected Forecast Errors for Combinations of  $\phi_1$  and  $\theta_1$ .

As the degree of forecast recovery worsens though, ARL and CDF performance decreases for all of the control charts. In regard to these traditional control charts, the CES chart provides the best compromising performance over a wide range of AR(1) process parameter values and shift sizes.

Performance of Control Charts Applied to ARMA(1,1) and MA(1) Models

For the ARMA(1,1) and MA(1) processes, the behavior of the forecast errors prior to the sustained level has an impact on all of the control charts.

Table 5: ARLs , MRLs, and CDFs for the ARMA(1,1) Process with Step Shift  $c = 1\sigma_\varepsilon$ .

ARMA Model	Control Chart	ARL	MRL	SRL	Cumulative Percentage of Signals Following Shift						
					1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>	6 <sup>th</sup>	7 <sup>th</sup>
1	IND	115	70	128	2.52	4.67	6.81	8.61	10.06	11.70	13.10
	EWMA	15	12	12	2.03	4.04	6.95	10.75	15.48	21.03	26.94
	CES	21	16	19	2.41	4.65	7.05	9.55	12.28	15.48	18.91
2	IND	279	191	284	2.52	3.23	3.62	3.98	4.28	4.62	5.00
	EWMA	136	96	141	7.15	8.32	9.35	10.05	10.74	11.31	11.91
	CES	184	136	177	3.03	3.88	4.41	4.86	5.27	5.66	6.13
3	IND	290	199	295	2.52	3.23	3.60	3.93	4.23	4.53	4.88
	EWMA	217	145	239	7.01	7.92	8.63	9.22	9.83	10.39	10.86
	CES	259	178	265	3.03	3.74	4.19	4.54	4.90	5.24	5.64
4	IND	270	177	293	2.52	4.34	5.97	7.14	8.10	8.93	9.75
	EWMA	227	150	252	7.03	8.16	8.96	9.48	10.16	10.71	11.22
	CES	252	165	276	3.03	4.77	6.32	7.39	8.32	9.08	9.86
5	IND	42	29	42	2.52	4.80	7.25	9.44	11.43	13.63	15.80
	EWMA	9	8	5	1.51	4.28	8.73	15.38	23.33	32.56	41.75
	CES	12	11	7	2.34	4.93	8.09	12.14	16.89	22.85	29.38
6	IND	177	122	180	2.52	2.94	3.68	4.13	4.59	5.10	5.67
	EWMA	35	30	26	2.59	3.38	4.34	5.13	6.09	7.28	8.46
	CES	51	43	39	2.45	2.95	3.71	4.30	4.95	5.68	6.53
7	IND	205	138	218	2.52	3.18	5.08	5.46	6.70	7.03	8.15
	EWMA	48	41	37	4.46	5.19	6.27	6.72	7.80	8.31	9.32
	CES	67	57	52	2.79	3.48	5.34	5.76	6.97	7.33	8.53
8	IND	64	43	63	2.52	3.38	5.84	6.85	8.78	9.92	11.85
	EWMA	13	11	8	1.61	3.06	5.99	9.03	14.30	19.20	26.22
	CES	17	15	11	2.36	3.40	6.11	7.77	11.08	13.76	18.36
AR(1) 9	IND	290	199	295	2.52	2.93	3.24	3.56	3.84	4.14	4.51
	EWMA	194	131	209	7.05	7.95	8.72	9.26	9.89	10.44	10.97
	CES	242	167	246	3.03	3.51	3.93	4.30	4.66	5.03	5.44
AR(1) 10	IND	119	81	120	2.52	3.27	4.08	4.91	5.60	6.41	7.21
	EWMA	21	19	14	2.24	3.51	4.82	6.24	8.10	10.16	12.67
	CES	29	26	20	2.44	3.28	4.18	5.22	6.32	7.59	8.98
MA(1) 11	IND	79	54	79	2.52	3.34	4.77	5.93	7.01	8.29	9.51
	EWMA	15	13	9	1.70	3.11	5.11	7.85	11.34	15.65	20.50
	CES	20	18	14	2.38	3.35	5.04	6.68	8.60	10.99	13.83
MA(1) 12	IND	117	78	120	2.52	2.93	5.13	5.49	7.04	7.45	8.92
	EWMA	22	19	14	2.24	3.05	4.63	5.43	7.64	9.02	11.96
	CES	29	26	20	2.44	2.91	4.99	5.45	7.32	8.01	10.07

Consider, for example, ARMA(1,1) Model 1 and AR(1) Model 10 in Table 5. Both exhibit a similar level of sustained forecast recovery (0.50 versus 0.53). Model 10's forecast errors attain a sustained level of shift at  $t = r + 1$ , while Model 1's forecast errors attain a sustained level at  $t = r + 34$ . Although Model 10 has a slightly higher sustained level of forecast recovery, the Individuals control chart performs better when applied to Model 1. The reason for this difference is a result of the magnitude of the gradually recovering forecast errors of Model 1. The Individuals chart takes advantage of the magnitude of forecast errors from time periods  $t = r + 1$  to  $t = r + 33$ . Again, the AR(1) process forecast errors recover immediately to the sustained level of the shift at time period  $t = r + 1$ . The other control charts also exhibit similar behavior when applied to these two models.

Consider another example using Model 1 compared with MA(1) Model 12 in Table 5. Both exhibit similar levels of sustained forecast recovery (0.50 versus 0.53), and both models attain a sustained level of shift at approximately  $t = r + 34$ . Both models exhibit gradually recovering forecast errors, but again the magnitude of the recovering forecast errors has a profound effect on the control charts. While the forecast errors arising from Model 1 gradually decrease from  $E(e_t^*) = 0.95$  to 0.50, those for Model 12 oscillate between values from  $E(e_t) = 0.10$  to 0.91 until converging upon the sustained level of the shift at  $E(e_t^*) = 0.53$ . As a result of this oscillating behavior, the control charts applied to Model 12 do not perform as well as the same charts applied to Model 1 even though Model 12 has a higher sustained level of the shift.

Many ARMA(1,1) processes exhibit oscillating behavior of forecast errors to some degree. The worst cases are those in which the forecast errors oscillate between values that alter in sign as well as magnitude and finally converge to the sustained level of the shift. ARMA(1,1) Models 3 and 4 in Table 5 are good examples of forecast errors exhibiting this oscillation behavior. Table 1 displays this behavior numerically for a shift of size  $c = 1\sigma_\varepsilon$ . The forecast errors in Model 3 oscillate between sequential values that differ in sign as well as absolute magnitude. The forecast errors in Model 4 oscillate between sequential

values that differ in sign, but the absolute magnitudes of the forecast errors are very similar.

The behavior of the forecast errors in Model 4 dampens the performance of any control chart that requires the summing or averaging of forecast errors over time such as the EWMA or CUSUM control charts. If the forecast errors differ in sign but not in absolute magnitude, the result is a canceling-out effect of summed or averaged forecast errors, until finally reaching the sustained level of the shift. Models producing forecast errors that differ in sign as well as absolute magnitude (Model 3) experience the same canceling out effect but not to the same degree as is seen in Model 4.

Consider a comparison of the performance of control charts applied to Models 3 and 4. Both exhibit the same level of sustained forecast recovery (0.03). Model 3's forecast errors attain a sustained level of shift at  $t = r + 5$ , while Model 4's forecast errors attain a sustained level at  $t = r + 38$ . Longer time until sustained recovery is attained usually provides for an all around better chart performance for a given sustained level of a shift, but the oscillation behavior of the forecast errors in Model 4 negates this advantage in the case of the EWMA control chart. The Individuals control chart takes advantage of the magnitude of the recovering forecast errors in Model 4, ignoring the sign of each forecast error value. As a result, the Individuals chart applied in Model 4 was found to have phenomenally better ARL, MRL, and CDF performance than in the case of Model 3, over all shift sizes. In contrast, the EWMA control chart applied in Model 4 was found to perform significantly worse than in the case of Model 3 providing ARLs, MRLs, and CDFs that are lower for every shift size. Although the EWMA chart suffers in Model 4, the good performance of the Individuals chart results in CES control chart performance that is also good.

#### Recommendations

As a result of the phenomenon of forecast recovery and the behavior of recovering forecast errors, the authors have several recommendations in regards to selecting the appropriate control chart to use with various autocorrelated processes. The practitioner should:

1. Determine the appropriate ARMA model and parameters regarding the process to be monitored, and use Eq. (9) to estimate the degree of forecast recovery.
2. Use Eq. (4) to determine the effect of forecast recovery on the forecast errors that will result from a step-shift of size  $c$  in the mean of the underlying ARMA process.
3. Use one of Eq. (5), (6), or (7), depending on if the model is an ARMA(1,1), AR(1), or MA(1), to determine the sustained level of recovery resulting from the step-shift of size  $c$ . The expected behavior of the recovering forecast errors should also be studied in regards to the rate of recovery, oscillation, the magnitude and sign of recovering forecast if oscillating, and the expected sampling period when the forecast will recover to the sustained level.
4. Select and apply the control chart who's performance is least affected by the forecast recovery, in face of the magnitude of the shift to be detected as well as the behavior of the recovering forecast.

The practitioner should take note that in the selection of the control chart, one first determines the magnitude of the shift that is deemed most important to detect. Recall, while the Individuals chart is best suited for rapidly detecting relatively large shifts, the EWMA chart is best suited for the eventual detection of small shifts. The CES chart serves as a compromise. Second, one must bear in mind that the behavior of recovering forecast might yield an otherwise favorable chart unsuitable for the monitoring the process at hand.

### Conclusion

This article provided a description of various models for autocorrelated data, as well as an introduction to the Box-Jenkins OSA forecast and forecast error often used to monitor an autocorrelated process. Also provided was a mathematical description of the impact of forecast recovery on the ARMA(p,q) process, and particularly the ARMA(1,1), AR(1), and MA(1) processes.

Additionally, the article included a discussion concerning the relationship between initial/sustained rates of forecast recovery, and a

model's particular parameter values and first order autocorrelation structure. It was shown that while the rates of forecast recovery differ for all models, these recovery rates are indeed a function of the model parameters. Additionally, knowledge of first order autocorrelation was shown helpful in determining the degree of sustained forecast error recovery in the ARMA(1,1), AR(1), and MA(1) processes. Examples were given of various ARMA(p,q) forecast error recovery rates over time, while tables were provided relating the sustained expected value of forecast errors for a wide variety of ARMA(p,q) processes.

Finally, it was found that the sustained level of forecast recovery following a shift had a tremendous effect on the performance of each control chart examined. The rate of recovery as well as the absolute magnitude and sign of forecast errors prior to attaining the sustained level of recovery were found to greatly influence the performance of the control charts. It was shown that for a given shift and sustained level of recovery, the control charts generally perform worse when applied to the forecast errors arising from AR(1) processes. The worsening of performance was shown to be due to the sudden forecast recovery characteristics inherent in these processes. As a result of the phenomenon of forecast recovery and the behavior of recovering forecasts, recommendations were made in regards to a practitioner selecting the most appropriate control chart for various ARMA processes.

### References

- Adams, B. M., Woodall, W. H., & Superville, C. R. (1994). Discussion of run-length distributions of special-cause control charts for correlated processes by D. G. Wardell, H. Moskowitz, and R. D. Plante. *Technometrics*, 36, 19-21.
- Alwan, L. C., & Roberts, H. V. (1988). Time-series modeling for statistical process control. *Journal of Business and Economic Statistics*, 6, 87-95.
- Atienza, O. O., Tang, L. L., & Ang, B. W. (1998). A SPC procedure for detecting level shift of autocorrelated processes. *Journal of Quality Technology*, 30, 340-351.



- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis, forecasting and control*. (3<sup>rd</sup> ed.). NJ: Prentice-Hall, Engelwood Cliffs.
- Lin, W. S. W., & Adams, B. M. (1996). Combined control charts for forecast-based monitoring schemes. *Journal of Quality Technology*, 28, 289-301.
- Lu, C. W., & Reynolds, M. R. Jr. (2001). CUSUM charts for monitoring an autocorrelated process. *Journal of Quality Technology*, 33, 316-334.
- Lu, C. W., & Reynolds, M. R. Jr. (1999a). EWMA control charts for monitoring the mean of autocorrelated processes. *Journal of Quality Technology*, 31, 166-188.
- Lu, C. W., & Reynolds, M. R. Jr. (1999b). Control charts for monitoring the mean and variance of autocorrelated processes. *Journal of Quality Technology*, 31, 259-274.
- Montgomery, D. C., & Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23, 179-204.
- Superville, C. R., & Adams, B. M. (1994). An evaluation of forecast-based quality control schemes. *Communications in Statistics: Simulation and Computation*, 23, 645-661.
- Vasilopoulos, A. V., & Stoamboulis, A.P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, 10, 20-30.
- Wardell, D. G., Moskowitz, H. and Plante, R. D. (1994). Run-length distributions of special-cause control charts for correlated processes. *Technometrics*, 36, 3-16.

## Determining Predictor Importance In Multiple Regression Under Varied Correlational And Distributional Conditions

Tiffany A. Whittaker  
University of Texas at Austin

Rachel T. Fouladi  
University of Texas  
M.D. Anderson Cancer Center  
at Houston

Natasha J. Williams  
University of Texas at Austin

---

This study examines the performance of eight methods of predictor importance under varied correlational and distributional conditions. The proportion of times a method correctly identified the dominant predictor was recorded. Results indicated that the new methods of importance proposed by Budescu (1993) and Johnson (2000) outperformed commonly used importance methods.

Key words: Multiple Regression; Predictor Importance; Relative Importance; Multicollinearity.

---

### Introduction

One of the most common statistical techniques used today is Multiple Regression (MR) Analysis (Neter, Kutner, Nachtsheim, & Wasserman, 1996). Once the predictors are selected for the MR model, researchers typically wish to establish the relative importance of the predictors when predicting the dependent variable. According to Healy (1990), the most typical request of statistical consultants when conducting MR analyses is to determine the relative importance of the predictor variables in the model, with the key focus on the question: Of all the predictors in the MR model, which one influences the criterion variable the most?

According to Kruskal (1984), there are two motives as to why relative importance is so

meaningful to researchers: 1) technological motives and 2) scientific motives. The technological motive is produced from the hopes of implementing change that is effective and economical. For example, “what should we attend to first in trying to reduce cancer deaths, improve education, maintain our systems of highways, increase productivity growth, etc.” (Kruskal, 1984, p. 39). The scientific motive is produced from the attempt to increase one’s basic understanding of some phenomenon with no concern of implementing immediate change. For example, “which variables should we examine in our next experiment or survey...since we never have the resources to examine all?” (Kruskal, 1984, p. 39). Regardless of the motive, predictor importance is of great concern when conducting MR analyses.

Consider  $p$  predictors,  $x_1 \dots x_p$ , of the criterion variable  $y$ . When the predictor variables in the MR model are perfectly uncorrelated, relative importance can simply be determined from the squared value of the zero-order correlations between the criterion and each of the predictors ( $\rho_{yx_j}^2$ ,  $j = 1 \dots p$ ) which, in that case, sum to the model’s squared multiple correlation (Budescu, 1993):

$$\rho_{y \cdot x_1 \dots x_p}^2 = \sum_{j=1}^p \rho_{yx_j}^2 \quad (1)$$

---

Tiffany A. Whittaker and Natasha J. Williams, Department of Educational Psychology, University of Texas at Austin. Rachel T. Fouladi, Department of Behavioral Science, University of Texas MD Anderson Cancer Center in Houston. Tiffany A. Whittaker’s email address: [Twhittaker@mail.utexas.edu](mailto:Twhittaker@mail.utexas.edu). Natasha J. Williams’ email address: [Tashwill@aol.com](mailto:Tashwill@aol.com). Rachel T. Fouladi’s email: [Rfouladi@mail.mdanderson.org](mailto:Rfouladi@mail.mdanderson.org).

Thus, the relative contribution of each predictor may be expressed in terms of percentages, as can be seen from the following equation (Lindeman, Merenda, & Gold, 1980, p. 119):

$$\text{Percentage Contribution} = 100 \frac{\rho_{yx_j}^2}{\rho_{y \cdot x_1 \dots x_p}^2}, \quad (2)$$

and this can be interpreted as the percentage of total variance in the criterion accounted for by a predictor. However, when the predictors are correlated with each other, which is normally the case, the above relationship is no longer viable. This is because part of a predictor's contribution becomes a shared contribution with one or more of the other predictor variables with which it happens to be correlated (Lindeman et al., 1980).

Many techniques have been proposed to assess the relative importance of predictors in ordinary least squares (OLS) MR models, with little consensus on which method is best employed (for reviews, see Budescu, 1993; Darlington, 1968). Proposed methods to determine the importance of the  $j$ th predictor of  $y$  include: 1) the squared zero-order correlation between the criterion variable and the predictor,  $\rho_{yx_j}^2$ ; 2) the standardized regression coefficient for the predictor in the  $p$ -predictor MR model,  $\beta_j^*$ ; 3) the  $t$ -statistic for the test of the regression coefficient in the  $p$ -predictor MR model,  $t_j$ ; 4) the product of the standardized regression coefficient for a predictor and its zero-order correlation with the criterion (Pratt, 1987),  $\beta_j^* \rho_{yx_j}$ ; 5) the squared partial correlation of the criterion variable and the predictor,  $\rho_{yx_j \cdot x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2$ ; and 6) the squared semi-partial correlation of the criterion variable and the predictor,  $\rho_{y(x_j \cdot x_1 \dots x_{j-1} x_{j+1} \dots x_p)}^2$  (c.f., Darlington, 1968; Budescu, 1993; Johnson, 2000). All of these methods of determining predictor importance provide the same information when the predictors are not intercorrelated. However, the information they provide is not equivalent when the predictors are correlated (Darlington, 1968).

The lack of consensus as to which importance method to use is understandable when one considers the differences between these methods, the most visible difference being the definition of importance adopted when using these

various methods (Budescu, 1993). For instance, the squared value of the zero-order correlation between the criterion and the predictor,  $\rho_{yx_j}^2$ , is the proportion of variance in the criterion accounted for by the predictor (Cohen & Cohen, 1975). Thus, it only illustrates a predictor's direct effect on the criterion (Budescu, 1993). Standardized regression coefficients,  $\beta_j^*$ , are interpreted as the amount of change that occurs in the criterion variable for each standard deviation change in a predictor variable while holding all other predictors in the model constant (Bring, 1994).

Hence, a predictor's importance is dependent upon its own contribution to the model, which is contingent upon the other predictors' contributions (Budescu, 1993). The  $t$ -values associated with the estimates of the coefficients for the predictors are computed to test the null hypothesis that each population regression coefficient in the model is equal to zero ( $\beta_j = 0$ ) (Lindeman et al., 1980). When computing a  $t$ -value for a predictor, it represents the increase in the model's squared multiple correlation when adding the predictor to the MR model after all the additional  $p - 1$  predictors have already been included in the MR model (Bring, 1994). Hence, a predictor's importance is dependent upon its own contribution to the model, which is contingent upon the other predictors' contributions. The product of the standardized regression coefficient for a predictor and its zero-order correlation with the criterion (Pratt, 1987),  $\beta_j^* \rho_{yx_j}$ , represents both a predictor's total effect ( $\beta_j^*$ ) and direct effect ( $\rho_{yx_j}$ ). The squared partial correlation,  $\rho_{yx_j \cdot x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2$ , and the predictor's "usefulness" (i.e., the squared semipartial correlation),  $\rho_{y(x_j \cdot x_1 \dots x_{j-1} x_{j+1} \dots x_p)}^2$ , (Darlington, 1968) can be perceived as the proportion of variance in the criterion that can be explained by each predictor variable contingent upon the other predictors' contributions (Budescu, 1993). Evidently, the definition of importance varies widely from method to method. Accordingly, these methods can often lead to different conclusions as to the relative importance of the same predictor variables (Budescu, 1993).

Dominance Analysis

Budescu (1993) recently suggested a new method, called Dominance Analysis, that identifies predictor importance while accounting for a predictor’s direct, partial, and total effect. Where  $x_i$  and  $x_j$  are a pair of predictors in the original set of  $p$  predictors, and  $x_h$  is any subset of the remaining  $p-2$  predictors,  $x_i$  “weakly dominates”  $x_j$ , if the following relationships among squared multiple correlations hold for all possible  $x_h$ :

$$\rho_{y \cdot x_i x_h}^2 \geq \rho_{y \cdot x_j x_h}^2 \tag{3}$$

or

$$(\rho_{y \cdot x_i x_h}^2 - \rho_{y \cdot x_h}^2) \geq (\rho_{y \cdot x_j x_h}^2 - \rho_{y \cdot x_h}^2), \tag{4}$$

where  $\rho_{y \cdot x_i x_h}^2$  is the squared multiple correlation of the model which includes predictor  $x_i$  and the remaining predictors,  $x_h$ , while excluding predictor  $x_j$ . After establishing pairwise “dominance or equality” for each  $p(p-1)/2$   $x_i x_j$  pairings, the next step is to compute

$$C_{x_i}^{(k)} = \sum (\rho_{y \cdot x_i x_h}^2 - \rho_{y \cdot x_h}^2) / m \tag{5}$$

for each variable  $x_i$  across all  $m$  models with  $k + 1$  predictors ( $x_i$  and  $k = 0 \dots p - 1$  variables), where  $x_h$  is any possible subset of  $k$  predictors with  $x_i$  excluded and  $m = \binom{p-1}{k}$ . Lastly, Budescu advises the computation of

$$C_{x_i} = \sum_{k=0}^{p-1} C_{x_i}^{(k)} / p, \tag{6}$$

which provides a meaningful decomposition of the  $p$ -predictor model’s squared multiple correlation.

Johnson’s Index

Johnson (2000) critiqued Budescu’s method and noted that computations are tedious and require more time as the number of predictor variables in the model increases (Johnson, 2000). Johnson (2000) suggested an alternative method that yields similar results with less computation, extending the work of Gibson (1962), Johnson

(1966), and Green Carroll, and DeSarbo (1978). Without loss of generality, let  $\mathbf{X}$  be an  $N \times p$  full-rank matrix of predictor scores in standard score form, and  $\mathbf{y}$  be the  $p \times 1$  criterion score vector also in standard score form. Singular value decomposition yields  $\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}'$ , where  $\mathbf{P}$  consists of eigenvectors of  $\mathbf{X}\mathbf{X}'$ ,  $\mathbf{Q}$  consists of eigenvectors of  $\mathbf{X}'\mathbf{X}$ , and  $\mathbf{\Delta}$  is the diagonal matrix with the square roots of corresponding eigenvalues on the diagonal. Let  $\mathbf{Z} = \mathbf{P}\mathbf{Q}'$ , which yields a best-fitting (minimum sum of squared residuals) set of orthogonal variables to  $\mathbf{X}$ . Let the regression of  $\mathbf{y}$  on  $\mathbf{Z}$  yield the vector of regression weights  $\boldsymbol{\beta}_Z^*$ , and the regression of  $\mathbf{X}$  on  $\mathbf{Z}$  yield the matrix of regression weights  $\mathbf{\Lambda}^*$ . Using the notation,

$$\mathbf{\Lambda}^{*[2]} = \|\lambda_{jk}^2\| \tag{7}$$

and

$$\boldsymbol{\beta}^{*[2]} = \|\beta_{z_{jk}}^{*2}\|, \tag{8}$$

Johnson’s index for each predictor’s relative importance is obtained from the elements of  $\boldsymbol{\epsilon} = \mathbf{\Lambda}^{*[2]} \boldsymbol{\beta}^{*[2]}$ , which when summed yield the original  $p$ -predictor model’s squared multiple correlation (Johnson, 2000).

Using an actual data set, Johnson compared his method ( $\epsilon$ ) with seven other measures of importance. These seven measures included the following: 1) the squared zero-order correlation between the criterion and the predictor; 2) the squared value of the standardized regression coefficient; 3) the product of the standardized regression coefficient for a predictor and its zero-order correlation with the criterion,  $\beta_j^* \rho_{yx}$ ; 4) the  $t$ -statistic associated with a predictor; 5) the squared value of the standardized partial regression coefficient from regressing the criterion on the orthogonal predictors (Gibson, 1962); 6) Green, Carroll, and DeSarbo’s (1978) relative weight measure ( $\delta_j^2$ ); and 7) Budescu’s (1993) Dominance Analysis method ( $C_{x_i}$ ). Relative weights for various predictor variables were calculated using each of the different importance methods. Johnson concluded that his method ( $\epsilon$ ),

Budescu's (1993) method ( $C_{x_i}$ ), and Green et al.'s (1978) method ( $\delta_j^2$ ) were comparable in terms of the relative weights assigned to the predictors and that these methods are the most efficient in obtaining the indirect and direct effects of the predictors on the criterion variable.

Johnson further examined the efficiency of his method by comparing it to both Budescu's (1993)  $C_{x_i}$  and Green et al.'s (1978)  $\delta_j^2$  across various regression models. Using 31 different sets of data (both authentic and simulated), Johnson calculated the relative importance weights assigned by each of the three different methods. The number of predictors in the MR model varied from 3 to 10, and the mean correlation among predictor variables varied from .10 to .70. Using Budescu's (1993) method as the standard, mean differences between the weights were calculated across the predictor variables. Johnson found that the mean difference between his method and Budescu's (1993) method was smaller than the mean difference between Budescu's method and Green et al.'s (1978) method. The mean differences between the relative importance weights were not related to the number of predictors in the model, but were related to the mean correlation among predictors in the model. Thus, Johnson's and Budescu's methods demonstrated similar findings as to the relative weights assigned, but as the mean correlation between the predictor variables increased, so did the differences between Johnson's and Budescu's (1993) methods. Still, as the mean correlation among predictors increased, Green et al.'s (1978) method deviated more from Budescu's (1993) method than Johnson's method. Johnson attributed the deviation between his method and Budescu's (1993) method to the fact that regression coefficients become unstable under conditions of multicollinearity, suggesting that both measures may generate questionable results under these conditions. Nevertheless, Johnson (2000) did not report which method performed the best in terms of correctly identifying the known dominant or most important predictor. In addition, results were not reported with respect to the performance of the predictor importance methods under various distributional conditions, such as multivariate nonnormality.

Normality of predictor and criterion variables is not an assumption of MR, however, nonnormality of predictor and criterion variables may create nonnormality in the error (residual) distributions, which is an assumption of MR. A violation of this assumption affects the validity of significance tests, such as  $t$ -tests, and increases the sample to sample variance of the regression coefficients. These effects are both due to the increase in the standard errors for the regression coefficients which occurs when the errors are nonnormally distributed (Hamilton, 1992).

Therefore, this study seeks to compare the performance of the new importance methods (i.e., Johnson's and Budescu's methods) to the other proposed measures of predictor importance in terms of identifying the known, correct dominant predictor. In addition, the current study will investigate the performance of these methods under a range of sample and distributional conditions using simulated data as well as a sample data set.

## Methodology

### Monte Carlo Study

A Monte Carlo simulation experiment was first conducted to compare methods of predictor importance under conditions of normality and nonnormality in the predictors and criterion, homogenous correlations among predictors, and heterogeneous correlations between predictors and the criterion. Data were generated from multivariate normal and nonnormal populations using the Headrick and Sawilowsky (1999) approach, which has been proposed as an alternative to other methods used for generating skewed and kurtotic distributions (e.g., Vale & Maurelli, 1983).

The correct identification of the known dominant predictor was examined under the following conditions:

Methods of Importance. Eight methods of importance were investigated. These included: 1) the squared zero-order correlation between the criterion variable and the predictor,  $\rho_{yx_j}^2$ ; 2) the standardized regression coefficient for the predictor in the  $p$ -predictor MR model,  $\beta_j^*$ ; 3) the  $t$ -statistic for the test of the regression coefficient in the  $p$ -predictor MR model,  $t_j$ ; 4) the product of

the standardized regression coefficient for a predictor and its zero-order correlation with the criterion (Pratt, 1987),  $\beta_j^* \rho_{yx_j}$ ; 5) the squared partial correlation of the criterion variable and the predictor,  $\rho_{yx_j \cdot x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2$ ; 6) the squared semi-partial correlation of the criterion variable and the predictor,  $\rho_{y(x_j \cdot x_1 \dots x_{j-1} x_{j+1} \dots x_p)}^2$ ; 7) Budescu's (1993) dominance measure,  $C_{x_j}$ , and 8) Johnson's (2000) Epsilon index,  $\epsilon_j$ .

Correlations among predictors. To represent low, moderate, and high multicollinearity levels among the predictor variables, data were generated from populations where predictors were homogeneously intercorrelated where the magnitude of the correlations equaled .10, .40, or .70.

Correlations between dominant predictor and criterion. Data were from populations where the predictors were heterogeneously correlated with the criterion. To establish known dominance of a predictor, the most important predictor correlated .40 or .60 with the criterion while the correlation between the additional predictors and the criterion equaled .30.

Distribution type. Data were distributed from both multivariate normal and nonnormal distributions, where the levels of skew and kurtosis for the predictors and the criterion were (sk, ku): (0, 0) for a normal distribution, (0, 6) for a symmetric and heavy-tailed distribution, or (2, 6) for an asymmetric and heavy-tailed distribution. These levels of skew and kurtosis were selected to compare the performance of the importance methods under the normal distribution as well as under some commonly encountered nonnormal distributions (Micceri, 1989).

Number of predictors,  $p$ . To represent a low, moderate, and high number of predictors in the MR model, data were from  $p$ -variate multinormal and multi-nonnormal populations, where  $p$  equaled 4, 6, or 8.

Sample size,  $n$ . To represent a wide range of sample sizes similar to those that may be encountered in the health, behavioral, and social sciences where extremely small as well as large sample studies are conducted, data were generated at specific ratios of sample size to number of variables, where  $n$  was either  $2p$ ,  $4p$ ,  $10p$ ,  $20p$ , or  $40p$ .

The six factors were fully crossed and each condition was replicated 1,000 times. Under each condition, the number of times that the correct predictor was identified as dominant was recorded.

## Results

A six-way factorial ANOVA [8 (methods of importance)  $\times$  3 (correlations among predictors)  $\times$  2 (correlations between dominant predictor and criterion)  $\times$  3 (distribution type)  $\times$  3 (number of predictors)  $\times$  5 (sample size)], with repeated measures on the importance methods, was performed on the hit rates. However, only a maximum of three-way interactions was investigated.

Four-way and five-way interactions were not investigated because separate ANOVAs for each importance method indicated that the three-way ANOVA models accounted for more than 90% of the variance in the hit rates ( $R^2$  ranged from .93 to .96). Because differential performance of the importance methods was the focus of the current research, only the interactions between the repeated measures factor (importance method) and the additional between-subjects factors were examined, as well as the main effect for importance method.

To control for Type I error, only those interactions with the repeated measures factor that obtained a significance level less than .001 were examined. These interactions consisted of the following and are discussed in this order: Importance Method  $\times$  Correlation Between Dominant Predictor and Criterion  $\times$  Sample Size; Importance Method  $\times$  Correlation Among Predictors  $\times$  Sample Size; Importance Method  $\times$  Correlation Among Predictors; Importance Method  $\times$  Sample Size. The Least Significant Difference (LSD) test was used for post hoc multiple comparisons. Again, to control for Type I error, only the pairwise differences that obtained a significance level less than .001 were examined.

Importance Method  $\times$  Correlation Between Dominant Predictor and Criterion  $\times$  Sample Size. The ANOVA indicated a significant interaction between importance method, correlation between dominant predictor and criterion, and sample size,  $F(28, 840) = 2.20$ ,  $p <$

.001 ( $\eta^2 = .07$ ). Post-hoc tests indicated that when the correlation between dominant predictor and criterion was low (.40) and sample size was small ( $2p$ ), Budescu's method and Johnson's  $\epsilon_j$  method performed comparably, outperforming the standardized regression coefficient and the method endorsed by Pratt (1987) (the product of the standardized regression coefficient for a predictor and its zero-order correlation with the criterion) in terms of identifying the dominant predictor (see Figure 1a); the standardized regression coefficient was outperformed by all of the other seven methods.

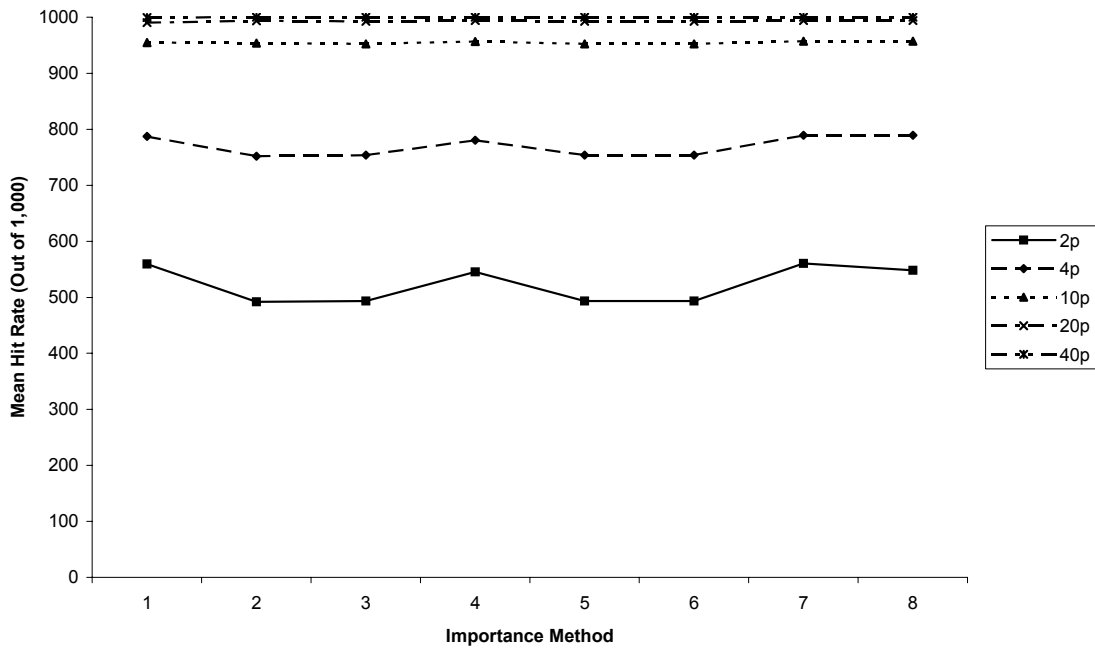
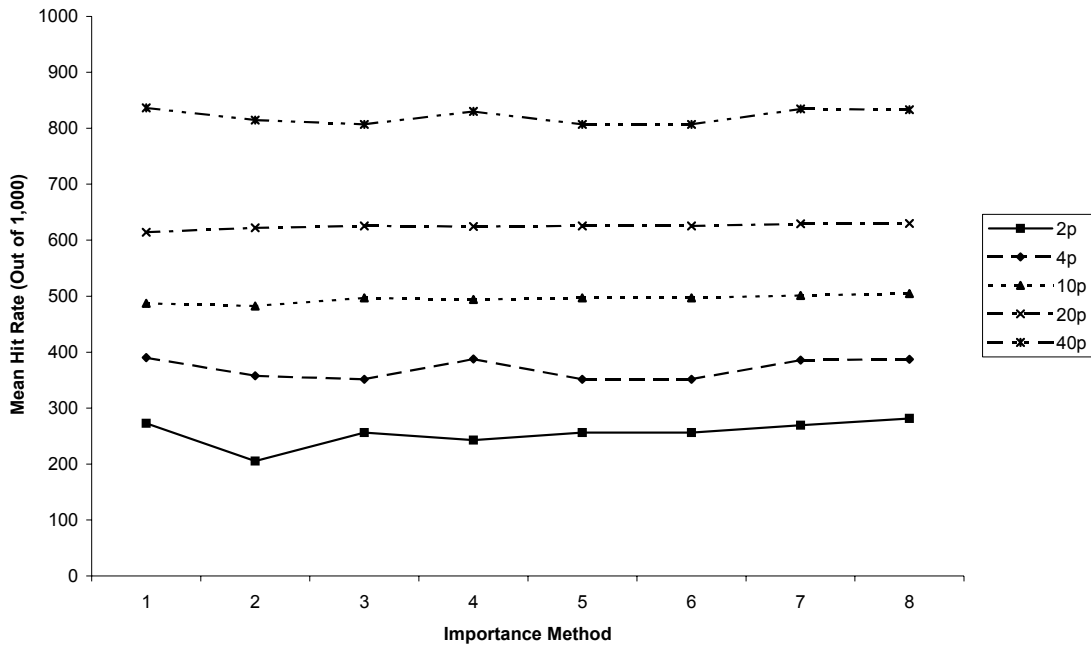
When the correlation between dominant predictor and criterion was low (.40) and sample size was at  $4p$ , Budescu's and Johnson's methods again performed comparably, outperforming the  $t$ -statistic, the squared partial correlation, and the squared semi-partial correlation; Pratt's method significantly outperformed the standardized regression coefficient while the squared zero-order correlation did not significantly differ from any of the other importance methods. There were no significant differences between the importance methods when sample sizes ranged from  $10p$  to  $40p$ .

When the correlation between the dominant predictor and criterion was high (.60) and sample size was low ( $2p$ ), the squared zero-order correlation, Pratt's method, Budescu's method, and Johnson's method all performed comparably and outperformed the standardized regression coefficient, the  $t$ -statistic, the squared partial correlation, and the squared semi-partial correlation (see Figure 1b). When the correlation between dominant predictor and criterion was high (.60) and sample size was at  $4p$ , Budescu's method and Johnson's method again performed comparably, outperforming the  $t$ -statistic, the squared partial correlation, and the squared semi-partial correlation while Budescu's and Pratt's

methods outperformed the standardized regression coefficient; the squared zero-order correlation did not significantly differ from any of the importance methods in terms of identifying the dominant predictor. There were no other significant differences between importance methods for sample sizes ranging from  $10p$  to  $40p$ .

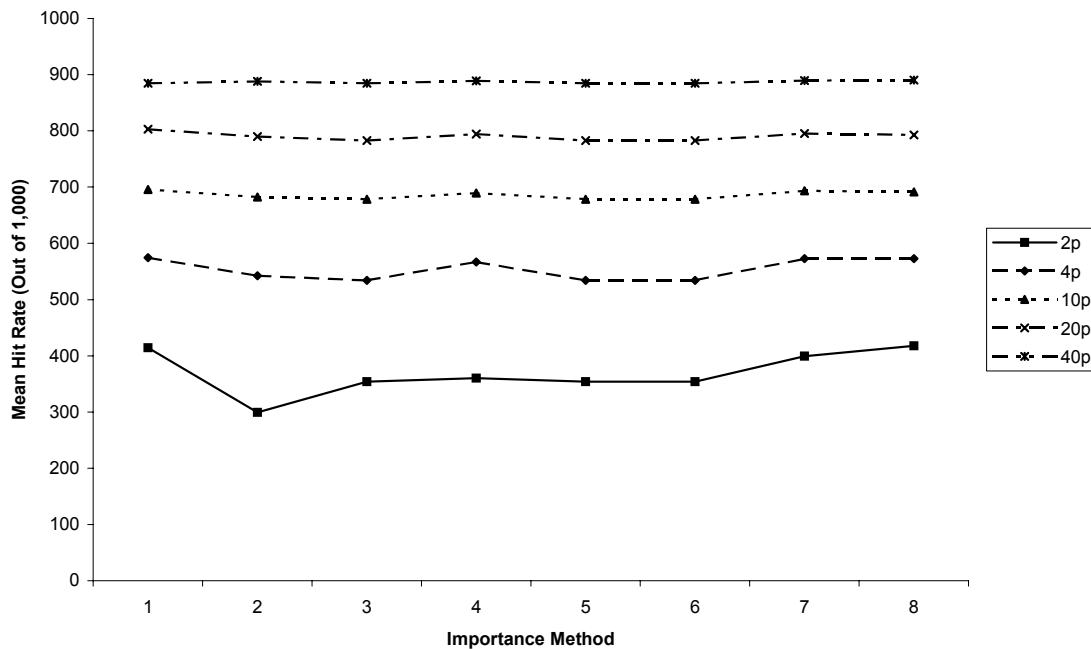
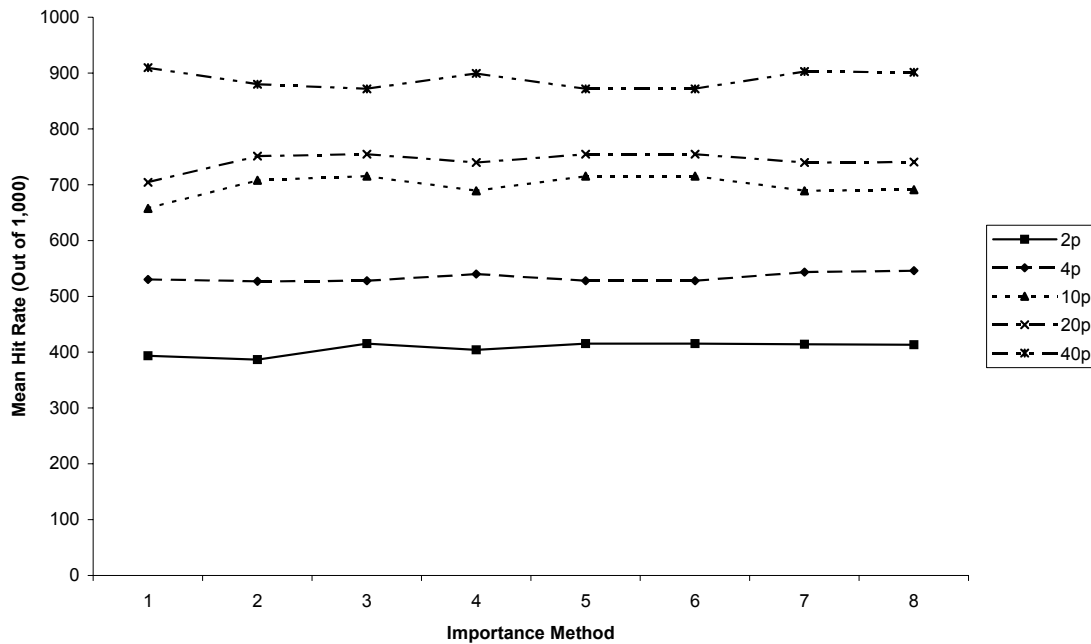
**Importance Method  $\times$  Sample Size.** The ANOVA also indicated a significant interaction between importance method and sample size,  $F(28, 840) = 4.84, p < .001$  ( $\eta^2 = .14$ ). Post hoc tests indicated that when sample size was small ( $2p$ ), the squared zero-order correlation, Budescu's method, and Johnson's method performed comparably, significantly outperforming the standardized regression coefficient, the  $t$ -statistic, Pratt's method, the squared partial correlation, and the squared semi-partial correlation (see Table 2); Pratt's method significantly outperformed the standardized regression coefficient.

When the sample size was  $4p$ , Pratt's method, Budescu's method, and Johnson's method performed comparably, significantly outperforming the standardized regression coefficient, the  $t$ -statistic, the squared partial correlation, and the squared semi-partial correlation; the squared zero-order correlation did not significantly differ from any of the other importance methods. No other significant differences were detected at other sample sizes ( $10p$ - $40p$ ).



Figures 1a-b. Mean hit rates (out of 1,000 replications) as a function of importance method and sample size at a) low (.40), and b) high (.60) correlation between dominant predictor and criterion. Importance methods are: 1 = squared zero-order correlation; 2 = standardized regression coefficient; 3 = *t*-statistic; 4 = Pratt's method; 5 = squared partial correlation; 6 = squared semi-partial correlation; 7 = Budescu's method; 8 = Johnson's method.





Figures 2a-b. Mean hit rates (out of 1,000 replications) as a function of importance method and sample size at a) low (.10), and b) moderate (.40) correlation among predictors. Importance methods are: 1 = squared zero-order correlation; 2 = standardized regression coefficient; 3 =  $t$ -statistic; 4 = Pratt's method; 5 = squared partial correlation; 6 = squared semi-partial correlation; 7 = Budescu's method; 8 = Johnson's method.

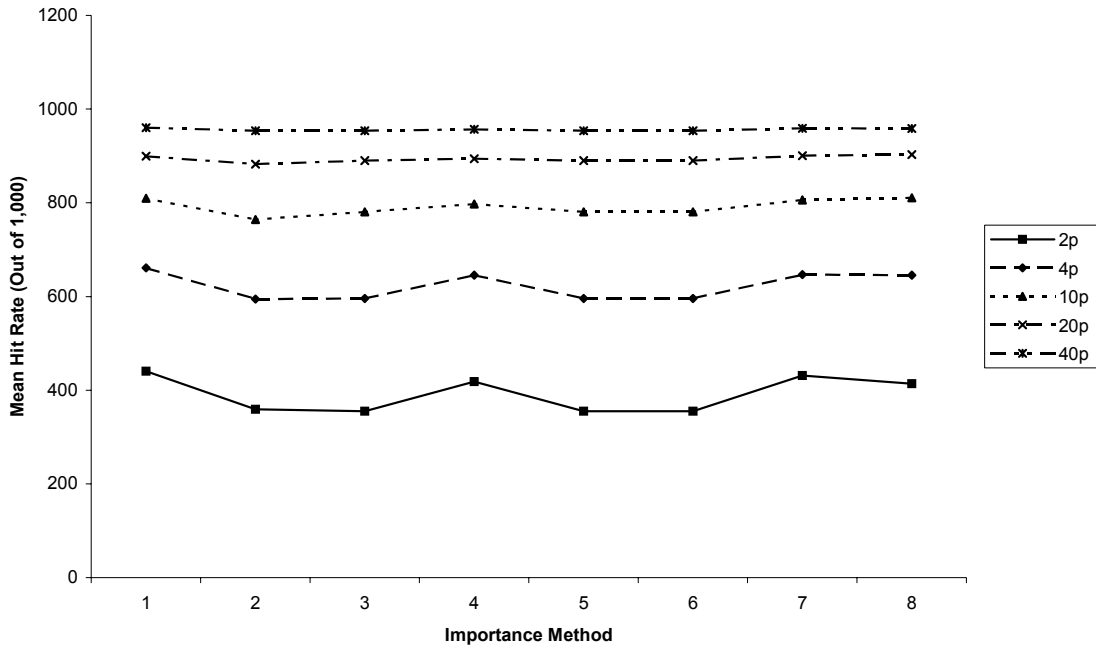


Figure 2c. Mean hit rates (out of 1,000 replications) as a function of importance method and sample size at high (.70) correlation among predictors. Importance methods are: 1 = squared zero-order correlation; 2 = standardized regression coefficient; 3 = *t*-statistic; 4 = Pratt’s method; 5 = squared partial correlation; 6 = squared semi-partial correlation; 7 = Budescu’s method; 8 = Johnson’s method.

Table 1 Mean Number of Hits (Standard Deviations) out of 1,000 as a Function of Correlation Among Predictor

	$\rho_{yx_j}^2$	$\beta_j^*$	$t_j$	$\beta_j^* \rho_{yx_j}$	$\rho_{yx_j \cdot x_1 \dots x_p}^2$	$\rho_{y(x_j \cdot x_1 \dots x_p)}^2$	$C_{x_j}$	$\epsilon_j$
Correlation Among Predictors								
.10	639.04 (308.56)	650.57 (280.38)	657.09 (276.92)	654.34 (288.04)	657.09 (276.92)	657.09 (276.92)	657.84 (289.53)	658.39 (287.38)
.40	674.47 (258.34)	640.32 (285.23)	646.70 (272.93)	659.93 (269.14)	646.70 (272.93)	646.70 (272.93)	669.93 (260.16)	672.99 (257.07)
.70	754.11 (258.55)	710.88 (290.35)	715.03 (286.93)	742.28 (266.94)	715.03 (286.93)	715.03 (286.93)	748.61 (264.19)	746.11 (268.64)

Main Effect of Importance Method. The ANOVA also indicated a significant main effect of importance method,  $F(7, 840) = 20.01, p < .001$  ( $\eta^2 = .14$ ). The mean number of hits out of 1,000 for each importance method is reported in Table 3. Post hoc tests indicated that Budescu’s method ( $C_{x_j}$ ), and Johnson’s index ( $\epsilon_j$ ) performed similarly by outperforming the remaining

measures when identifying the dominant predictor, with the exception of the squared zero-order correlation. The squared zero-order correlation and Pratt’s method significantly outperformed the standardized regression coefficient, the *t*-statistic, the squared partial correlation, and the squared semi-partial correlation, which all performed comparably.

Table 2: Mean Number of Hits (Standard Deviations) out of 1,000 as a Function of Sample Size

Sample Size	$\rho_{yx_j}^2$	$\beta_j^*$	$t_j$	$\beta_j^* \rho_{yx_j}$	$\rho_{yx_j \cdot x_1 \dots x_p}^2$	$\rho_{y(x_j \cdot x_1 \dots x_p)}^2$	$C_{x_j}$	$\varepsilon_j$
2p	416.04 (195.50)	348.50 (182.51)	374.91 (177.38)	394.09 (186.74)	374.91 (177.38)	374.91 (177.38)	414.80 (188.25)	414.70 (186.70)
4p	588.69 (254.93)	554.56 (250.73)	552.59 (254.55)	584.02 (249.80)	552.59 (254.55)	552.59 (254.55)	587.48 (255.49)	588.11 (254.33)
10p	720.98 (269.44)	718.04 (252.66)	724.56 (246.10)	725.30 (256.93)	724.56 (246.10)	724.56 (246.10)	729.39 (254.59)	731.04 (253.97)
20p	802.15 (234.72)	807.96 (208.31)	809.19 (208.01)	809.24 (215.61)	809.19 (208.01)	809.19 (208.01)	811.85 (215.50)	812.06 (215.47)
40p	918.19 (106.50)	907.22 (117.46)	903.46 (124.86)	914.94 (105.94)	903.46 (124.86)	903.64 (124.86)	917.13 (103.89)	916.57 (104.46)

Table 3: Mean Number of Hits (Standard Deviations) out of 1,000

$\rho_{yx_j}^2$	$\beta_j^*$	$t_j$	$\beta_j^* \rho_{yx_j}$	$\rho_{yx_j \cdot x_1 \dots x_p}^2$	$\rho_{y(x_j \cdot x_1 \dots x_p)}^2$	$C_{x_j}$	$\varepsilon_j$
689.21 <sup>ab</sup> (279.33)	667.26 <sup>c</sup> (285.99)	672.94 <sup>bc</sup> (279.58)	685.52 <sup>bd</sup> (276.79)	672.94 <sup>bc</sup> (279.58)	672.94 <sup>bc</sup> (279.58)	692.13 <sup>a</sup> (273.58)	692.50 <sup>ad</sup> (273.03)

Note. Means that share the same letter superscript do not significantly differ.

### Conclusion

One of the primary reasons for conducting this study was to determine which importance measure performs better in terms of identifying the correct dominant predictor. Similar to Johnson's (2000) findings, this Monte Carlo study indicates that Budescu's method ( $C_{x_j}$ ) and Johnson's index ( $\varepsilon_j$ ) perform comparably in terms of identifying the dominant predictor. Overall, both Budescu's and Johnson's methods also outperform the additional importance methods, with the exception of the squared zero-order correlation.

Trends did appear in the interactions that further substantiate the use of either Budescu's method or Johnson's method when determining predictor importance, especially under very small sample size conditions (2p-4p). As sample size increased (at 10p), however, the differences between all the importance methods became negligible, regardless of multicollinearity or dominance level. Budescu's method did differ from Johnson's method under the various levels of multicollinearity, in that Johnson's method performed better than Budescu's under moderate

multicollinearity with a very small sample size (2p), whereas Budescu's method performed better than Johnson's under high multicollinearity with a very small sample size (2p). Again, however, as sample size increased, the differences between these two methods became negligible under these multicollinearity conditions. The squared zero-order correlation did not appear to differentiate itself as a viable measure of importance as it did not significantly differ from additional importance methods under certain conditions.

Interestingly, two of the factors investigated in the current study did not interact with the various importance methods in either two-way or three-way interactions, such as the number of predictors in the MR model or distribution type. This indicates that no significant differences emerge between the importance methods as a function of the levels of either of these factors. Still, the levels of the factors used in the current study may not have been extreme enough to be able to examine differences between importance methods. Thus, future studies could examine the effect of MR models with a larger

number of predictors under more extreme levels of multivariate nonnormality.

In the current study, the *t*-statistic, the squared partial correlation, and the squared semi-partial correlation all performed identically, identifying the dominant predictor the same number of times under each condition. This may have been due to the homogeneous correlations among the predictor variables. As a result, real and simulated data sets with heterogeneous correlations among predictors were used to determine if these methods would differ under such conditions. The results of these analyses indicated that these three methods still identified the dominant predictor identically, indicating that the similarities between these three methods must be due to their definitions. In other words, all three methods are related to the variance in the model's multiple squared correlation that is attributable to a predictor variable after consideration of the additional variables' contribution to the model's squared multiple correlation.

#### Nursing Facility Consumer Satisfaction Survey

In an effort to improve the quality of care provided in nursing facilities, the Nursing Facility Consumer Satisfaction Survey (NFCSS) was developed (c.f., Cortés, Montgomery, Morrow, & Monroe, 2000). The survey consists of 12 items that assess general and specific consumer satisfaction with nursing facility care in certain domains, such as incontinence, physical activity, and medication management. Two versions of the survey were developed, one for nursing home residents and the other for family respondents. Each item is scored using a 7-point Likert scale ranging from 1 (very dissatisfied) to 7 (very satisfied).

In the first phase of a statewide longitudinal study, the survey was administered to a total of 138 family respondents of residents across 100 nursing facilities (Fouladi, 2001). For the purposes of this paper, 3 items which assess different types of activity satisfaction were selected to predict general satisfaction with the goal of identifying which activity satisfaction item is most associated with general satisfaction. One predictor variable was represented by the item on the survey: "How satisfied are you with the facility's ability to provide activities that your family member enjoy(s)?", to which responses

symbolized satisfaction with enjoyable or recreational activities.

The second predictor variable was represented by the item: "How satisfied are you with the facility's ability to provide activities that keep your family member as physically active as possible?", which symbolized satisfaction with physical activities. The third predictor was represented by the item: "How satisfied are you with the facility's ability to provide activities that keep your family member as mentally alert as possible?", which symbolized satisfaction with mental alertness activities. The criterion variable represented overall satisfaction with the nursing facility and corresponded to the item: "Overall, how satisfied are you with your family member's experience in this nursing facility?".

These four items on the survey are shown in the Appendix. This particular model was selected due to the high level of multicollinearity among the predictor variables and the moderate correlation between each predictor variable and the criterion. In addition, the distributional properties of the variables in the data set are comparable to the distributional properties of the variables from the simulation study. Intercorrelations among the predictor variables and the criterion variable and their descriptive statistics are shown in Table 4.

#### Results

Table 5 shows the predictor variables' relative weights assigned by each importance method. With the exception of the squared zero order correlation and Pratt's method,  $\beta_j^* \rho_{yx_j}$ , all of the importance methods selected the enjoyable activities predictor (predictor 1) as the most important variable. In contrast, the squared zero order correlation selected the physical activities predictor as most important and Pratt's method,  $\beta_j^* \rho_{yx_j}$ , assigned the same weights to both enjoyable and physical activities, producing a tie between these two variables in terms of importance.

#### Conclusion

This data set demonstrates how similar both Budescu's ( $C_{x_j}$ ) and Johnson's ( $\epsilon_j$ ) methods are in

that they assigned identical weights to each predictor variable. Excluding the squared zero order correlation and Pratt's method,  $\beta_j^* \rho_{yx_j}$ , all of the importance methods performed similarly to these two new methods, selecting enjoyable activities as the most important of the three predictor variables. Nonetheless, these additional methods do not take into account a predictor's direct and indirect effects as do both Budescu's ( $C_{x_j}$ ) and Johnson's ( $\epsilon_j$ ) methods.

Researchers typically wish to establish the relative importance of predictors in MR models. Many techniques are used to do this, however, no consensus exists as to which is best. This is due to the common problem of multicollinearity, which renders the typical methods ambiguous and

dependent upon the measure's definition of importance.

Budescu (1993) and Johnson (2000) have both established methods of importance that attempt to control for multicollinearity problems. The results of the simulation study are consistent with Johnson's (2000) finding that Budescu's method and Johnson's index perform comparably.

However, Budescu's method requires one to perform all possible regressions, which becomes fatiguing as the number of predictors in the MR model increases. Because Budescu's measure and Johnson's index performed comparably, it appears that Johnson's index would be the most computationally efficient measure to use if one is interested in determining predictor importance while accounting for a predictor's

Table 4: Nursing Facility Consumer Satisfaction Survey Variables' Intercorrelations and Descriptive Statistics (N = 138)

<i>Variables</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
1. Enjoyable Activities	--	.63*	.59*	.49*
2. Physical Activities		--	.73*	.50*
3. Mental Alertness Activities			--	.45*
4. Overall Satisfaction				--
Mean	6.01	5.77	5.79	6.25
Standard Deviation	1.02	1.19	1.10	0.93
Skew	-1.93	-1.72	-1.42	-1.85
Kurtosis	5.55	3.57	2.64	4.29

Note. \*  $p < .001$ .

Table 5  
*Comparison of Relative Weights Calculated by Each Importance Method for the NFCSS Data*

<i>Predictors</i>	$\rho_{yx_j}^2$	$\beta_j^*$	$t_j$	$\beta_j^* \rho_{yx_j}$	$\rho_{yx_j \cdot x_1 \dots x_p}^2$	$\rho_{y(x_j \cdot x_1 \dots x_p)}^2$	$C_{x_j}$	$\epsilon_j$
Enjoyable Activities	.24	.27	2.80	.13	.06	.04	.12	.12
Physical Activities	.25	.25	2.26	.13	.04	.03	.11	.11
Mental Alertness Activities	.20	.11	.99	.05	.01	.01	.08	.08

Note.  $N = 138$ . Average intercorrelation (in absolute value) among predictors = .65.

direct and total effects.

Future research should examine how various importance methods perform with heterogeneous correlations among predictor variables, which is typically the case with MR

models. The focus of the current study was to determine the correct known dominant predictor, which is a commonly asked question by researchers. Still, there are instances in which researchers wish to know the rank order of

predictor importance. In other words, which is the most important, the next most important, etc. Thus, future research could be implemented to investigate the performance of importance methods in terms of identifying the correct ranking of predictor variable importance.

The effects of multicollinearity and multivariate nonnormality on the importance methods were of particular interest in the current study. Although multicollinearity did affect the performance of relative importance methods, multivariate nonnormality did not. This is encouraging because multivariate nonnormality is typically found in real world data sets (Micceri, 1989). Additional research could examine extreme levels of multivariate nonnormality to determine whether there is a threshold at which point nonnormality does affect importance methods.

#### References

- Bring, J. (1994). How to standardize regression coefficients. *The American Statistician*, *48*(3), 209-213.
- Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*(3), 542-551.
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cortés, L. L., Montgomery, E. W., Morrow, K. A., & Monroe, D. M. (December, 2000). *A statewide assessment of quality of care, quality of life and consumer satisfaction in Texas Medicaid nursing facilities*. Austin, TX: Texas Department of Human Services, Long Term Care Office of Programs, Medical Quality Assurance.
- Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*(3), 161-182.
- Fouladi, R. T. (January, 2001). *Texas Department of Human Services Rider 26 longitudinal study of nursing home quality of care*. Austin, TX: Texas Department of Human Services, Long Term Care Office of Programs, Medical Quality Assurance.
- Gibson, W. A. (1962). Orthogonal predictors: A possible resolution of the Hoffman-Ward controversy. *Psychological Reports*, *11*, 32-34.
- Green, P. E., Carroll, J. D., & DeSarbo, W. S. (1978). A new measure of predictor importance in multiple regression. *Journal of Marketing Research*, *15*, 356-360.
- Hamilton, L. C. (1992). *Regression with graphics*. Belmont, CA: Duxbury Press.
- Headrick, T. C. & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distribution: Extending the Fleishman power method. *Psychometrika*, *64*(1), 25-35.
- Healy, M. J. R. (1990). Measuring importance. *Statistics in medicine*, *9*, 633-637.
- Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, *35*(1), 1-19.
- Johnson, R. M. (1966). The minimal transformation to orthonormality. *Psychometrika*, *31*, 61-66.
- Kruskal, W. (1984). Concepts of relative importance. *Questiio*, *8*(1), 39-45.
- Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott, Foresman and Company.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1990). *Applied linear statistical models: Regression, analysis of variance, and experimental designs*. (3<sup>rd</sup> ed.). Chicago, IL: Richard D. Irwin, Inc.
- Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In T. Pukkila & S. Puntanen (Eds.), *Proceedings of Second International Tampere Conference in Statistics* (pp. 245-260). University of Tampere, Finland.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465-471.

## Robust Estimation Of Multivariate Failure Data With Time-Modulated Frailty

Pingfu Fu  
Epidemiology & Biostatistics  
Case Western Reserve University

J. Sunil Rao  
Epidemiology & Biostatistics  
Case Western Reserve University

Jiming Jiang  
Department of Statistics  
University of California

---

A time-modulated frailty model is proposed for analyzing multivariate failure data. The effect of frailties, which may not be constant over time, is discussed. We assume a parametric model for the baseline hazard, but avoid the parametric assumption for the frailty distribution. The well-known connection between survival times and Poisson regression model is used. The parameters of interest are estimated by generalized estimating equations (GEE) or by penalized GEE. Simulation studies show that the procedure is successful to detect the effect of time-modulated frailty. The method is also applied to a placebo controlled randomized clinical trial of gamma interferon, a study of chronic granulomatous disease (CGD).

Key words: Frailty models; multivariate failure data; generalized linear models.

---

### Introduction

In the analysis of failure time data, one of the common assumptions made is that the life histories for subjects under study are statistically independent (at least conditionally on the observed fixed-time covariates). This assumption may be violated when individuals within some subgroup (e.g. siblings or parents in the same family, litter mates in animal study) share common unmeasured factors. Frailty models have been widely used for correlated survival data after Vaupel et. al. (1979) introduced the concept of frailty for making adjustments for the over-dispersion (heterogeneity) in their mortality study.

A frailty is an unobserved random effect shared by subjects within a subgroup. These include shared frailty (Hougaard, 1986a), bivariate frailty (Xue, 1998) as well as correlated frailty (Yashin, et. al. 1995), but few of them deal with time-dependent frailty (Self, 1995; Yau and McGilchrist, 1998). Most papers in the literature assume that individuals in the same cluster are

born at a certain level of relative frailty and stay at this level through out life. As mentioned by Vaupel et. al. (1979), this may not be true in reality, for example, in human population mortality study, the frailty of an individual is large during an early period of life, after which it stabilizes, followed by an increasing frailty due to the natural aging process. For univariate frailty model, there are several limitations, for example, the model only allows positive correlations within the cluster, and the unobserved factor (frailty) is the same within the cluster (Xue, 1998).

Typically we assume that the frailty acts multiplicatively on each individual's hazard rate. We propose a time-modulated frailty model to analyze multivariate failure time data. The proposed model is more general than other frailty models, having as special members regular frailty models, such as shared frailty and bivariate frailty models if we ignore the time-modulated component in the model. Using the well-known connection to Poisson regression (Aitkin and Clayton, 1980), the derived model is a generalized linear mixed model (glmm). We adopt a robust approach for estimating some parameters using the generalized estimating equations (GEE) in this Poisson regression setting. For other parameters, the estimating procedures are equivalent to a generalized penalized estimating equations

---

Pingfu Fu is senior instructor, email address: [pxfl6@po.cwru.edu](mailto:pxfl6@po.cwru.edu). This is a portion of Fu's PhD dissertation. J. Sunil Rao is Associate professor, email address: [sunil@hal.cwru.edu](mailto:sunil@hal.cwru.edu). Jiming Jiang is Associate professor, e-mail address: [jiang@wald.ucdavis.edu](mailto:jiang@wald.ucdavis.edu).

(GPPE). Under this approach, we do not specify the exact distribution of frailty and in this sense, our approach is robust.

Model construction

Self (1995) introduced a time-dependent frailty model

$$\lambda_i(t) = Y_i(t)\zeta_i(t)\lambda_0(t)\exp(\beta'x_i(t)),$$

where  $Y_i(t)$  and  $x_i(t)$  are predictable scalar and p-vector value processes, respectively,  $\zeta_i(t)$  is a stationary stochastic process with positive, continuous sample paths,  $\beta = (\beta_1, \beta_2, \dots, \beta_p)$  and  $\lambda_0(t)$  are unknown parameters. Instead of putting a stochastic process  $\zeta_i(t)$ , a time-dependent frailty process, in the hazard function, we introduce an "interaction" term between the frailty and time as a time-modulated frailty. In the following sections, we will give the model formulation in two different settings.

Single-level of clustering

The most common situation in the multivariate survival data is the time to the recurrence of some chronic disease for a patient, for example, breast cancer, or survival of litters of rats, survival of twins, etc. All these can be thought to consist of single-level clustering of data. The survival times in each cluster (patient, litter, twins) are correlated and the survival times between the clusters are assumed independent. Let the triple  $(T_{ik}, \delta_{ik}, x_{ik})$  represent the data, where  $i$  is the cluster index ( $i = 1, \dots, n$ ) consisting of correlated survival times  $T_{ik}$  ( $k = 1, \dots, n_i$ ). Thus, the  $k$ th individual in the  $i$ th group is modeled as

$$\lambda_{ik}(t) = w_i(t)\lambda_0(t)\exp(\beta'x_{ik}),$$

where  $w_i(t) = t^\theta \xi_i$  and  $\theta$  is unknown parameter. Here  $\xi_i$  are realizations of a nonnegative random variable with density function  $g(\xi)$ .

Assume  $E(\xi_i) = 1$  (see Nielsen et. al., 1992) and  $\text{var}(\xi_i) = \sigma^2$  for the distribution of the frailty  $\xi_i$ . When  $\theta = 0$ , the model is a shared frailty model,  $\xi_i \lambda_0(t) \exp(\beta'x_{ik})$ . The above model can also be easily generalized to the correlated

individual frailty model studied by Yashin et. al. (1995) by specifying  $w_i(t) = \xi_i + t^\theta \eta_i$  and letting  $n_i = 2$  and  $\theta = 0$ .

Multiple-levels of clustering

In some studies it may be reasonable to expect more than one level of within-cluster association. For example, the association between a parent and child versus that two siblings in studies of familial disease aggregation, or the durations inside and outside of hospitals for a patient who is admitted into a hospital several times for the same disease (Xue, 1998). The single-level clustering model can be extended to allow for grouping defined by multiple nested factors.

Again, suppose the data consists of the usual triple  $(T_{ijk}, \delta_{ijk}, x_{ijk})$ , using  $i$  to index the clusters (litters, families) ( $i = 1, 2, \dots, n$ ). Each cluster contains two distinguishable subgroups ( $j = 1, 2$ ). Within each cluster, individuals have correlated survival times  $T_{ijk}$  for  $k = 1, \dots, n_{ij}$ . When  $n_{ij} = 1$ , then  $(T_{i11}, T_{i21})$  is bivariate survival time, for example, as used in the adult Danish twins study (Hougaard et. al., 1992). We will assume the frailty acts multiplicatively on the individual's hazard with following form

$$\lambda_{ijk}(t) = w_{ij}(t)\lambda_0(t)\exp(\beta'x_{ijk}),$$

where  $w_{ij}(t) = t^\theta \eta_{ij}$  and  $\eta_{i1}, \eta_{i2}$  are the realizations of two correlated random variables with nonnegative values (with joint density function  $h(u, v)$ ). The  $\eta_{ij}$  is the frailty for the  $i$ th cluster and  $j$ th subgroup. The frailties can be characterized by a parametric bivariate distribution, for example,

$$(\log(\eta_1), \log(\eta_2)) \sim N(0,0; \sigma_1^2, \sigma_2^2, \sigma_{12}).$$

We also assume  $E(\eta_{ij}) = 1, i = 1, \dots, n, j = 1,2, \text{var}(\eta_{ij}) = \sigma_j^2$  and  $\text{cov}(\eta_1, \eta_2) = \sigma_1 \sigma_2 \rho$ . If  $\theta = 0$ , then it is a case studied by Xue (1998); if  $\theta > 0$  or  $\theta < 0$ , then we can see that the effect of frailty increases or decreases as time increases.

As we can see from the model construction in both single-level and multiple-level of clustering cases, given the frailty, its effect on the hazard changes over time.



For the exponential model, the baseline cumulative hazard is  $\Lambda_0(t) = t$ , and the hazard function becomes  $\lambda_{ijk}(t | w_{ij}) = t^\theta \eta_{ij} \exp(\beta'x_{ijk})$ .

For the Weibull model,  $\Lambda_0(t) = t^\nu$ , and the hazard function is

$$\lambda_{ijk}(t | w_{ij}) = t^\theta \eta_{ij} \nu t^{\nu-1} \exp(\beta'x_{ijk}).$$

We assume that observations between different clusters are independent and given the frailty  $w_{ij}$  (namely  $\eta_{i1}$  and  $\eta_{i2}$ ), the observations in each cluster are conditionally independent. It can be shown that, approximately,

$$\delta_{ijk} | (\eta_{i1}, \eta_{i2}) \sim \text{Poisson}(\mu_{ijk}),$$

where

$$\mu_{ijk}(t) = e^{\beta'x_{ijk}} \int_0^t w_{ij}(u) \lambda_0(u) du.$$

The details are given in Appendix 1.

**Robust estimation procedures**

As described in Appendix 1, we can treat the censoring variable as a correlated Poisson random variable with degree of over-dispersion depending on its mean. Since the full likelihood method is not feasible without numerical integration, and because of the intractability of the marginal likelihood function, we may apply the generalized estimating equations (GEE) approach (Liang and Zeger, 1986), which only requires the specification of the first two moments of the responses for each individual.

As mentioned by Hougaard (1984), the choice of the frailty distribution is crucial since the results for the survival population will be rather different with different frailties. In the following section, we will examine this robust approach, which only requires up to second-order of moments of the frailty distribution. It is robust in the sense that the full likelihood is not required and a fully parametric assumption for the frailty is avoided. The following procedures are for the single-level of clustering case, but they can be easily generalized to the multiple-level clustering case.

**Exponential case**

**Estimation of coefficients**

We assume that the baseline hazard is from exponential distribution. Given the frailty  $\xi_i$  as mentioned before,  $\delta_{ik} | \xi_i \sim \text{Poisson}(\tilde{\mu}_{ik} \xi_i)$ ,

where  $\tilde{\mu}_{ik} = e^{\beta'x_{ik}} \frac{t_{ik}^{\theta+1}}{\theta + 1}$ . It is easy to get

following quantities from the formulae for the multiple-level of clustering case (see Appendix 1).

$$E(\delta_{ik}) = \tilde{\mu}_{ik} = e^{\beta'x_{ik}} \frac{t_{ik}^{\theta+1}}{\theta + 1},$$

$$\text{var}(\delta_{ik}) = E(\tilde{\mu}_{ik} \xi_i) + \text{var}(\tilde{\mu}_{ik} \xi_i) = \tilde{\mu}_{ik} + \tilde{\mu}_{ik}^2 \sigma^2$$

and the unconditional covariance

$$\begin{aligned} \text{COV}(\delta_{ik}, \delta_{il}) &= \text{COV}(\tilde{\mu}_{ik} \xi_i, \tilde{\mu}_{il} \xi_i) \\ &= \tilde{\mu}_{ik} \tilde{\mu}_{il} \sigma^2, k \neq l, \end{aligned}$$

$$\text{cov}(\delta_{ik}, \delta_{i'l'}) = \text{cov}(\tilde{\mu}_{ik} \xi_i, \tilde{\mu}_{i'l'} \xi_{i'}) = 0, i \neq i'.$$

In order to get the estimates of the regression parameters, we apply the quasi-likelihood score equations in spirit of GEE, i.e.

$$U_\beta(\beta, \theta) = \sum_{i=1}^n \left( \frac{\partial \tilde{\mu}_i}{\partial \beta} \right)' \text{Var}(Y_i)^{-1} (Y_i - \tilde{\mu}_i) = 0, \tag{1}$$

where  $Y_i' = (\delta_{i1}, \dots, \delta_{in_i})$  and  $\tilde{\mu}_i' = (\tilde{\mu}_{i1}, \dots, \tilde{\mu}_{in_i})$ .

Note that  $\text{Var}(Y_i) = \text{Var}(Y_i; \beta, \theta)$ , which depends on  $\beta$  and  $\theta$  in the above equations. Thus, we need the estimating procedure for  $\theta$ .

**Estimation of time-modulated frailty parameters**

The estimate of the variance component  $\sigma^2$  is treated as nuisance parameter, which is estimated by a method of moments defined as

$$\hat{\sigma}^2 = \frac{\sum_{i, n_i > 1} \sum_{k \neq k'} (\delta_{ik} - \hat{\mu}_{ik}) (\delta_{ik'} - \hat{\mu}_{ik'}) + \sum_{i, n_i = 1} [(\delta_{i1} - \hat{\mu}_{i1})^2 - \hat{\mu}_{i1}]}{\sum_{i, n_i > 1} \hat{\mu}_{ik} \hat{\mu}_{ik'} + \sum_{i, n_i = 1} \hat{\mu}_{i1}^2}.$$

The conditional likelihood function has form:

$$L_{ik}(\beta | \xi_i) = (\xi_i \tilde{\mu}_{ik})^{\delta_{ik}} e^{-\xi_i \tilde{\mu}_{ik}} \left[ \frac{w_i(t_{ik}) \lambda_0(t_{ik})}{\int_0^{t_{ik}} w_i(u) \lambda_0(u) du} \right]^{\delta_{ik}}$$

the second term in the above equation equals to  $\left[ \frac{(\theta + 1)}{t_{ik}} \right]^{\delta_{ik}}$ . Thus, the log of the likelihood function can be approximated as

$$l \approx l_Q(\beta, \theta) + \sum_{i,k} \delta_{ik} [\log(\theta + 1) - \log(t_{ik})], \quad (2)$$

where  $l_Q(\beta, \theta)$  is the log of the quasi-likelihood function for correlated Poisson variates.

We then introduce the penalized score equation for the  $\theta$ ,

$$\sum_{i=1}^n \left( \frac{\partial \tilde{\mu}_i}{\partial \theta} \right)' Var(Y_i)^{-1} (Y_i - \tilde{\mu}_i) + \kappa(n) \sum_{i,k} \frac{\delta_{ik}}{\theta + 1} = 0, \quad (3)$$

the equation (3) can be viewed as a regularized generalized estimating equation with a penalty

term  $\kappa(n) \sum_{i,k} \frac{\delta_{ik}}{\theta + 1}$ , where  $\kappa(n) = n^{-\tau}$  with

$\tau > 0$ . When the tuning parameter  $\kappa(n) = 1$ , the left hand side of equation (3) is the partial derivative  $\frac{\partial l}{\partial \theta}$ . The estimators for  $\beta$  and  $\theta$

can be obtained by iterating between (1) and (3).

### Weibull case

When the baseline hazard is assumed to have a Weibull distribution, the model is more flexible by introducing an additional scale parameter  $\nu$ .

### Estimation of coefficients

As before, given the frailty,

$$\delta_{ik} | \xi_i \sim \text{Poisson}(\tilde{\mu}_{ik} \xi_i),$$

where  $\tilde{\mu}_{ik} = e^{\beta x_{ik}} \frac{\nu}{\theta + \nu} t_{ik}^{\theta + \nu}$ . Similarly, we have

$$E(\delta_{ik}) = \tilde{\mu}_{ik} = e^{\beta x_{ik}} \frac{\nu}{\theta + \nu} t_{ik}^{\theta + \nu},$$

$$\text{var}(\delta_{ik}) = E(\tilde{\mu}_{ik} \xi_i) + \text{var}(\tilde{\mu}_{ik} \xi_i) = \tilde{\mu}_{ik} + \tilde{\mu}_{ik}^2 \sigma^2$$

and the unconditional covariance

$$\text{COV}(\delta_{ik}, \delta_{il}) = \text{COV}(\tilde{\mu}_{ik} \xi_i, \tilde{\mu}_{il} \xi_i)$$

$$= \tilde{\mu}_{ik} \tilde{\mu}_{il} \sigma^2, k \neq l,$$

$$\text{cov}(\delta_{ik}, \delta_{i'l'}) = \text{cov}(\tilde{\mu}_{ik} \xi_i, \tilde{\mu}_{i'l'} \xi_{i'}) = 0, i \neq i'.$$

The estimate of the regression parameters can be obtained by the following generalized estimating equations, i.e.

$$U_\beta(\beta, \theta, \nu) = \sum_{i=1}^n \left( \frac{\partial \tilde{\mu}_i}{\partial \beta} \right)' Var(Y_i)^{-1} (Y_i - \tilde{\mu}_i) = 0, \quad (4)$$

where  $Y_i' = (\delta_{i1}, \dots, \delta_{in_i})$  and  $\tilde{\mu}_i' = (\tilde{\mu}_{i1}, \dots, \tilde{\mu}_{in_i})$ .

Note that  $Var(Y_i) = Var(Y_i; \beta, \theta, \nu)$  which depends on  $\beta, \theta$  and  $\nu$  in the above equations, as mentioned in exponential case, we have to get the  $n^{1/2}$ -consistent estimates for  $\nu$  and  $\theta$ .

### Estimation of other parameters

The estimate of the variance component  $\sigma^2$  is defined the same way as the exponential case:

$$\hat{\sigma}^2 = \frac{\sum_{i,\eta>1} \sum_{k \neq k'} (\delta_{ik} - \hat{\mu}_{ik}) (\delta_{ik'} - \hat{\mu}_{ik'}) + \sum_{i,\eta=1} [(\delta_{i1} - \hat{\mu}_{i1})^2 - \hat{\mu}_{i1}]}{\sum_{i,\eta>1} \hat{\mu}_{ik} \hat{\mu}_{ik'} + \sum_{i,\eta=1} \hat{\mu}_{i1}^2}.$$

The conditional likelihood function in this case has a form

$$L_{ik}(\beta | \xi_i) =$$

$$(\xi_i \tilde{\mu}_{ik})^{\delta_{ik}} e^{-\xi_i \tilde{\mu}_{ik}} \left[ \frac{w_i(t_{ik}) \lambda_0(t_{ik})}{\int_0^{t_{ik}} w_i(u) \lambda_0(u) du} \right]^{\delta_{ik}}$$

with the second term in the above equation equals to  $[\frac{(\theta + \nu)}{t_{ik}}]^{\delta_{ik}}$ . Thus, the log of the likelihood function can be approximated as

$$l \approx l_Q(\beta, \theta, \nu) + \sum_{i,k} \delta_{ik} [\log(\theta + \nu) - \log(t_{ik})], \tag{5}$$

where  $l_Q(\beta, \theta, \nu)$  is the log of the quasi-likelihood function for correlated Poisson variates. If we re-parameterized  $\theta + \nu$  as  $\varphi$ , then

$$\begin{aligned} \frac{\partial \tilde{\mu}_{ik}}{\partial \varphi} &= e^{\beta'x_{ik}} \left[ -\frac{\nu}{\varphi^2} t_{ik}^\varphi + \frac{\nu}{\varphi} t_{ik}^\varphi \log(t_{ik}) \right] \\ &= \tilde{\mu}_{ik} \left[ \log(t_{ik}) - \frac{1}{\varphi} \right], \end{aligned}$$

and

$$\frac{\partial \tilde{\mu}_{ik}}{\partial \nu} = e^{\beta'x_{ik}} \frac{t_{ik}^\varphi}{\varphi} = \frac{\tilde{\mu}_{ik}}{\nu}$$

Thus, we introduce the penalized score equations for  $\varphi$  as we did in the exponential case,

$$\begin{aligned} U_\varphi &= \sum_{i=1}^n \left( \frac{\partial \tilde{\mu}_i}{\partial \varphi} \right)' \text{Var}(Y_i)^{-1} (Y_i - \tilde{\mu}_i) \\ &+ \kappa(n) \sum_{i,k} \frac{\delta_{ik}}{\varphi} = 0, \end{aligned} \tag{6}$$

where the tuning parameter  $\kappa(n) = n^{-\tau}$ ,  $\tau > 0$  and when  $\tau = 0$ , the left hand side of equation (6) is  $\frac{\partial l}{\partial \varphi}$ . Because  $\nu$  is unidentifiable

from the score equations, we use plug-in estimate for it. Notice that, if we have estimates of  $\varphi$  and  $\beta$ , then, from equation  $\tilde{\mu}_{ik} = e^{\beta'x_{ik}} \frac{\nu}{\varphi} t_{ik}^\varphi$ . we can obtain the estimate of  $\nu$  by following formula,

$$\hat{\nu} = \frac{1}{N} \sum_{i,k} \frac{\tilde{\mu}_{ik} \varphi}{e^{\beta'x_{ik}} t_{ik}^\varphi}, \tag{7}$$

which is moment estimate if we replace  $\tilde{\mu}_{ik}$  by its sample mean.

In summary, we propose following algorithm for the estimates of  $\beta, \varphi, \nu$  and  $\theta$ ,

1. Given initial values of  $\varphi, \nu$ :  $\varphi^{(0)}, \nu^{(0)}$ , and fit Poisson regression by generalized estimating equations (4) using log link function with offset equals to

$$\log\left(\frac{\nu^{(0)}}{\varphi^{(0)}} t_{ik}^{\varphi^{(0)}}\right), \text{ and obtain } \beta^{(0)}$$

and get (update)  $\hat{\mu}_{ik}^{(0)}$ .

2. Update  $\varphi^{(0)}$  from equation (6).
3. Update  $\nu$  by following formula:

$$\hat{\nu}^{(1)} = \frac{1}{N} \sum_{i,k} \frac{\hat{\mu}_{ik}^{(0)} \varphi^{(1)}}{e^{\beta^{(0)'x_{ik}} t_{ik}^{\varphi^{(1)}}}.$$

4. Go to step 1, 2, and 3 again until the convergence criteria is satisfied.

Because  $\hat{\theta}$  is consistent and  $\text{var}_J(\hat{\theta})$  is asymptotically unbiased (see the results in Appendix 2 and 3), we can use statistic  $\frac{\hat{\theta} - \theta}{[\text{var}_J(\hat{\theta})]^{1/2}}$ , which is asymptotically  $N(0,1)$  for

inference; thus, the null hypothesis  $\theta = 0$  can be tested. If we reject the null hypothesis from the test, then we claim that the effect of time-modulated frailty exists. In the following sections, we examine our method by simulation followed by analyzing CGD dataset.

### Simulations

There is a difficulty with conducting simulations in this setting, since it's difficult to generate correlated survival times with time-modulated frailties as we can see it in the specification of the hazard function which involves time-modulated frailties.

We generate datasets of correlated Weibull (without time-modulated frailty, i.e.  $\theta = 0$ ) by using positive mixing distributions (Hougaard, 1986a) along with the random effects approach. Let  $T_{ik}$  be the survival times of observation  $k$  of individual (cluster)  $i$  conditional on an observed covariate  $Z_i$ . In this setup we

assume that the  $T_{ik}$ 's in different clusters are independent. Now assume  $Z$  to be positive stable with index  $\alpha$ . The Laplace transform for  $Z$  is  $E(\exp(-sZ)) = \exp(-s^\alpha)$ . If we now define another random variable  $Y_{ik}$  to be Weibully distributed with scale parameter  $\exp(\beta'x_{ik})$  and shape parameter  $a$ , then  $T_{ik} = Y_{ik} Z_i^{-1/a}$ . Thus the  $T_{ik}$ 's within a cluster are multivariate Weibull with Weibull margins having scale  $\exp(\alpha\beta'x_{ik})$  and shape  $\alpha a$ . The correlation between  $\log(T_{ik})$  and  $\log(T_{il})$  is then just  $1 - \alpha^2$  for  $k \neq l$ . The generation of positive stable variates  $Z_i$  can be done using Splus which employs Chambers et. al.'s (1976) algorithm.

Instead of choosing different values of index of positive stable random variable, different cluster size and different percentage of censoring, we just generate two datasets with clusters 50 and 150. In each cluster, there are 5 observations and the index of positive stable random variable  $\tilde{\alpha} = 0.6$ , the coefficient of the linear predictor  $\tilde{\beta} = 3$  and the shape parameter of the Weibull  $\tilde{\nu} = 2$ , thus, the marginal distribution of the correlated Weibull is still Weibull with shape parameter  $\nu = 1.2$  and the scale parameter  $\beta = 1.8$  (actually  $\exp(x'\beta)$ ) where  $x$  is from the design matrix which is 1 or 0 depending whether a random number from standard normal is nonnegative or negative. The survival times are censored at fixed value to achieve 10% censoring. The estimates of parameters interested are the means of 100 replicates. The tuning parameter in the penalized score equation is  $\kappa(n) = 1$  and  $\kappa(n) = n^{-1/30}$  which is arbitrarily picked. We understand that the optimal choice of the tuning parameter may be selected by many methods, for example, the cross validation approach.

In this correlated Weibull case, as we know, there is no time-modulated frailty in it. We still assume the time-modulated frailty model, and the frailty term is in the form of  $w_{ik}(t) = t_{ik}^\theta \eta_i$ , and  $\lambda_0(t) = \nu^{-1}$  is the baseline hazard from the Weibull distribution.

As we can see from Table 1, the parameter estimate of  $\theta$  is not significant from 0 for two

different values of tuning parameter which means we can not reject null hypothesis  $\theta = 0$  based on asymptotic Wald type test. Thus, there does not appear to be a time-modulated frailty effect in this dataset. The estimates of  $\beta$  and  $\nu$  are very close to the true values.

Table 1: Results of fitting the correlated Weibull by time -dependent frailty model with two values of  $\kappa(n)$  in the penalized score equation, number of clusters = 50 and 100 simulations.

Parameter	BC (no BC)		GJ Standard	
	Estimate	error	t Value	Pr >  t
-----				
$\kappa(n) = 1 :$				
$\beta$	1.783 ( 1.827)	0.2718	6.560	< 0.0001
$\theta$	0.001 (-0.097)	0.3463	0.001	0.9998
$\nu$	1.208 ( 1.316)	0.4406	2.742	0.0061
$\varphi$	1.208 ( 1.219)	0.1094	11.042	< 0.0001
-----				
$\kappa(n) = n^{-1/30} :$				
$\beta$	1.645 ( 1.696)	0.2578	6.381	< 0.0001
$\theta$	0.150 ( 0.093)	0.1704	0.879	0.3793
$\nu$	0.936 ( 1.016)	0.2605	3.593	0.00033
$\varphi$	1.086 ( 1.109)	0.1012	10.73	< 0.0001
-----				
$\beta$ (SN, 1993)	1.781	0.3852	4.624	< 0.0001

Note: The true value of  $\beta$  is 1.8 and 1.2 for  $\nu$ . BC stands for bias corrected, GJ for grouped jackknife, and SN for Segal and Neuhaus.

The estimate of  $\beta$  by our procedure is consistent with other two approaches. From the variance estimates of  $\beta$ , there is small gain in term of efficiency although there is no time-modulated frailty effect in this case.

The biased estimates (values in the 'no BC' column) overestimated the parameters when the tuning parameter  $\kappa(n) = 1$ , and underestimated when  $\kappa(n) = n^{-1/30}$ . The optimal tuning parameter  $\tau$  may be a positive value that is very close to zero. We can do further simulation for large number of clusters and for different values of  $\tau$ , as well as other parameters, such as different percentage of censoring, different value of index in the positive stable distribution. The results from Table 2 are more close to the true values, this is

because we have larger number of clusters (150 clusters) and the estimates of  $\beta$ ,  $\varphi$  and  $\theta$  are consistent.

Table 2: Results of fitting the correlated Weibull by time-dependent frailty model with two values of  $\kappa(n)$  in the penalized score equation, number of clusters = 150 and 100 simulations.

	BC (no BC)	GJ Standard		
Parameter	Estimate	error	t Value	Pr >  t
<hr/>				
$\kappa(n)=1$ :				
$\beta$	1.808 (1.822)	0.1526	11.85	< 0.0001
$\theta$	0.000 (-0.013)	0.1166	0.004	0.9968
$\nu$	1.205 (1.215)	0.1783	6.758	0.0001
$\varphi$	1.205 (1.203)	0.0644	18.71	< 0.0001
<hr/>				
$\kappa(n)=n^{-1/200}$ :				
$\beta$	1.779 (1.792)	0.1508	11.8	< 0.0001
$\theta$	0.034 (0.014)	0.1017	0.332	0.7396
$\nu$	1.147 (1.171)	0.1626	7.054	< 0.0001
$\varphi$	1.181 (1.185)	0.0635	18.6	< 0.0001
<hr/>				
$\beta$ (SN)	1.781	0.3852	4.624	< 0.0001

Note: The true value of  $\beta$  is 1.8 and 1.2 for  $\nu$ . BC stands for bias corrected, GJ for grouped jackknife, and SN for Segal and Neuhaus.

A real data example

The well-known Chronic Granulomatous Disease (CGD) dataset, which is described in the Appendix D of the book by Fleming and Harrington (1991), has been analyzed by many authors. CGD is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections, which usually present early life and may lead to death in childhood. Phagocytes from CGD patients ingest microorganisms normally but fail to kill them, primarily due to the inability to generate a respiratory burst dependent on the production of superoxide and other toxic oxygen metabolites. Thus, it is the failure to generate microbicidal oxygen metabolites within the phagocytes of CGD patients.

There is evidence that gamma interferon is an important macrophage activating factor which could restore superoxide anion production and

bacterial killing by phagocytes in CGD patients. In order to study the ability of gamma interferon to reduce the rate of serious infections, a double-blinded clinical trial was conducted in which patients were randomized to placebo vs. gamma interferon. The data we use here, which is a little different from the one used by Fleming and Harrington (1991) in the example at page 162, has 65 patients in placebo group, 63 in gamma interferon group, of 30 placebo patients who experienced at least one infection, 4 experienced 2, 4 experienced 3, 1 experienced 4, 1 experienced 5 and 1 experienced 7; of 14 treatment patients who experienced at least one infection, 4 experienced 2 and 1 experienced 3.

It is reasonable to assume that the patients' frailties are time-modulated, since the risk of infection may increase once a first failure event occurs. In this data set, we treat each patient as a cluster, and the frailty term is in the form of  $w_{ik}(t) = t_{ik}^{\theta} \xi_i$ .

Table 3. Results of fitting the CGD dataset by proposed method with other two models.

	BC (no BC)	GJ Standard		
Parameter	Estimate	error	t Value	Pr >  t
<hr/>				
$\kappa(n)=1$ :				
$\beta$	-0.835 (-0.856)	0.2588	-3.207	0.0013
$\theta$	1.293 (1.321)	0.1995	6.481	< 0.0001
$\varphi$	1.328 (1.357)	0.1945	6.828	< 0.0001
$\nu$	0.035 (0.037)	0.0184	1.944	0.052
<hr/>				
$\kappa(n)=n^{-1/30}$ :				
$\beta$	-0.822 (-0.845)	0.2468	-3.332	0.0009
$\theta$	1.116 (1.169)	0.1809	6.169	< 0.0001
$\varphi$	1.148 (1.204)	0.1736	6.613	< 0.0001
$\nu$	0.032 (0.034)	0.01461	2.204	0.0275
$\beta$ (SN, 1993)	-0.856	0.2489	-3.4389	0.00058

Note: BC stands for bias corrected, GJ for grouped jackknife, and SN for Segal and Neuhaus.

Table 3 provides estimates of  $\beta$  with several methods, the estimates of other parameters followed by standard error for case of  $\kappa(n) = 1$  by our time-modulated frailty model are  $\hat{\nu} = 0.035$  (0.0184),  $\hat{\theta} = 1.293$  (0.1995),  $\hat{\varphi} = 1.328$  (0.1945).

The negative value of  $\hat{\beta} = -0.8353$  means that the treatment (gamma interferon) effectively reduces the recurrence of pyogenic infections as compare to the placebo. The estimate of  $\beta$  is consistent to that from other approaches.

From the estimates of  $\theta$  and its variance, we can see that there is a time-modulated frailty effect in this dataset as noticed by Self (1995) though we have different model formulations. The parameter estimate of the time-modulated frailty  $\hat{\theta} = 1.293$  is statistically significant from 0; the positive sign also means that given the frailty, its effect on the hazard is increasing as the life goes on.

The estimate of the treatment effect  $\beta$  is consistent with other two approaches; all of them indicate a statistically significant difference between the gamma interferon and placebo. The time-modulated frailty model does not seem to improve the efficiency, but the proposed model does help us to understand the nature of the frailty. In CGD case, the existence of effect of time-modulated frailty means that if a patient has a large frailty at the beginning, then (s)he will have an increasing chance of recurrence of pyogenic infections.

### Conclusion

Few results about time-modulated frailty models are available in the literature (Yau and McGilchrist, 1998; Self, 1995). Our model provides one way to detect whether there is a trend in the hazard function with time given the frailty. Our model is different from Yau and McGilchrist's (1998), which assumes a different frailty for each time period of recurrence of disease; and different from Self's (1995) which introduces a stochastic process of frailty in the hazard function. The models proposed can also be extended in more general case, for example, in the multiple-level of clustering case, the time-modulated frailty can have the following form  $w_{ij}(t) = \xi_i + t^\theta \eta_{ij}$ ,

where  $\xi_i, (\eta_{i1}, \eta_{i2})$  are independent realizations of two independent random variables with positive values. The resulting models are more complex than the one we proposed. To fit this model, we may use techniques of nonlinear mixed-effects models (Pinheiro and Bates, 2000).

Clinically speaking, the significance of the model is to realize whether there is an effect of time-modulated frailty in some diseases. If it does exist, for example, the pyogenic infection case (CGD data), it will tell us that more frail patients (say, have recurrence at the beginning) are more likely to have recurrence late in their life, which may suggest that those patients need more aggressive treatment (e.g. high dosage).

### References

- Aitkin, M., & Clayton, D. G. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, 29, 156-163.
- Chambers, J. M., & Stuck, B. W. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association*, 71, 340-344.
- Clayton, D. G. (1983). Fitting a general family of failure-time distributions using GLIM. *Applied Statistics*, 32, 102-109.
- Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*, Clarendon: Oxford University Press.
- Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. New York: John Wiley & Sons.
- Hougaard P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71, 75-83.
- Hougaard P. (1986a). A class of multivariate failure time distributions. *Biometrika*, 73, 671-678.
- Hougaard P. (1986b). Survival models for heterogeneous population derived from stable distributions. *Biometrika*, 73, 387-396.
- Hougaard, P., Harvald, B., & Holm, N. V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930. *Journal of the American Statistical Association*, 87, 17-24.
- Jiang, J., & Zhang, W. (in press). Robust estimation in generalized linear mixed models. *Biometrika*.
- Laird, N., & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76, 3231-240.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. (2<sup>nd</sup> ed.). London: Chapman and Hall.

McGilchrist, C. A., & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, 47, 461-466.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sorensen, T. I. A.. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19, 25-43.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and Splus*. New York: Springer-Verlag.

Segal, M. R., & Neuhaus, J. M. (1993). Robust inference for multivariate survival data. *Statistics in Medicine*, 12, 1019-1031.

Self, S. (1995). A regression model for counting processes with a time dependent frailty. Technical report. Seattle, WA: Fred Hutchinson Cancer Center.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

Therneau, T. M., & Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, 16, 2029-2047.

Vaupel, J. W., Manton, K.G., & Stallard E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439-454.

Xue, X. (1998). Multivariate survival data under bivariate frailty: An estimating equation approach. *Biometrics*, 54, 1631-1637.

Yashin, A. I., Vaupel, J. W., & Iachine, I. A. (1995). Correlated individual frailty: An advanced approach to survival analysis of bivariate data. *Mathematical Population Studies*, 5, 145-159.

Yau, K. K. W., & McGilchrist, C. A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine*, 17, 1201-1213.

## Appendix 1: Likelihood and moments

Likelihood construction. For the model with multiple-levels of clustering, the hazard function is

$$\lambda_{ijk}(t) = w_{ij}(t)e^{\beta'x_{ijk}}\lambda_0(t).$$

Its corresponding density and survival functions:

$$f_{ijk}(t | w_{ij}) =$$

$$w_{ij}(t)\lambda_0(t)\exp(\beta'x_{ijk})\exp(-e^{\beta'x_{ijk}}\int_0^t w_{ij}(u)\lambda_0(u)du),$$

and

$$S_{ijk}(t | w_{ij}) = \exp(-e^{\beta'x_{ijk}}\int_0^t w_{ij}(u)\lambda_0(u)du).$$

Thus, the contribution of the  $i$ th individual to the conditional likelihood given frailty  $w_{ij}$  is

$$\begin{aligned} L_{ijk}(\theta, \beta | w_{ij}) &= f_{ijk}^{\delta_{ijk}}(S_{ijk})^{(1-\delta_{ijk})} \\ &= [w_{ij}(t)\lambda_0(t)\exp(\beta'x_{ijk})]^{s_{ijk}} \exp(-e^{\beta'x_{ijk}} \\ &\quad \int_0^t w_{ij}(u)\lambda_0(u)du) \\ &= [e^{\beta'x_{ijk}}\int_0^t w_{ij}(u)\lambda_0(u)du]^{s_{ijk}} \exp(-e^{\beta'x_{ijk}} \\ &\quad \int_0^t w_{ij}(u)\lambda_0(u)du) \left[ \frac{w_{ij}(t)\lambda_0(t)}{\int_0^t w_{ij}(u)\lambda_0(u)du} \right]^{\delta_{ijk}} \\ &= [\mu_{ijk}^{\delta_{ijk}} e^{-\mu_{ijk}}] \left[ \frac{w_{ij}(t)\lambda_0(t)}{\int_0^t w_{ij}(u)\lambda_0(u)du} \right]^{\delta_{ijk}}, \end{aligned}$$

where  $\mu_{ijk} = e^{\beta'x_{ijk}}\int_0^t w_{ij}(u)\lambda_0(u)du$  and

$w_{ij} = t^\theta \eta_{ij}$ . Because  $T_{ijk}$  are conditionally independent given  $w_{ij}$ , therefore the conditional likelihood is

$$\begin{aligned} L(\theta, \beta | \eta_1, \eta_2) \\ &= \prod \{ \mu_{ijk}^{\delta_{ijk}} e^{-\mu_{ijk}} \left[ \frac{w_{ij}(t_{ijk})\lambda_0(t_{ijk})}{\int_0^{t_{ijk}} w_{ij}(u)\lambda_0(u)du} \right]^{\delta_{ijk}} \}, \end{aligned}$$

and the conditional log likelihood:

$$\log(L) = \sum \{ \delta_{ijk} \log(\mu_{ijk}) - \mu_{ijk} + \delta_{ijk} [ \log(w_{ijk}(t_{ijk})) + \log(\lambda_0(t_{ijk})) - \log(\int_0^{t_{ijk}} w_{ij}(u) \lambda_0(u) du) ] \}.$$

Therefore, from the above arguments, we have the following :

**Result:** Given the frailties  $\eta_{i1}$  and  $\eta_{i2}$ ,  $\delta_{ijk}$  can be thought as a Poisson random variable with mean  $\mu_{ijk}$ . We will focus on the baseline hazard from Weibull distribution since it has a fairly flexible hazard function; baseline hazards from other distributions can be modeled by piece-wise exponential distribution which is a special case of Weibull. Assume the hazard function from Weibull distribution is  $\lambda(t) = \phi v^{v-1}$ , here  $\phi$  is a scale parameter, and  $v$  is a shape parameter. The Weibull distribution is flexible enough to accommodate increasing ( $v > 1$ ), decreasing ( $v < 1$ ) or constant hazard rate ( $v = 1$ ). When we have Weibull baseline distribution, the above log likelihood becomes

$$\log(L) = \sum \{ \delta_{ijk} \log(\frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v} e^{\beta x_{ijk}}) - \frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v} e^{\beta x_{ijk}} + \delta_{ijk} [ \log(t_{ijk}^\theta \eta_{ij}) + \log(v) - \log(\frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + 1}) ] \}.$$

Since for Weibull distribution, the baseline hazard

$$\text{is } \lambda_0(t) = v t^{v-1} \text{ and } \mu_{ijk} = e^{\beta x_{ijk}} \frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v},$$

$$\text{because } \int_0^{t_{ijk}} w_{ij}(u) \lambda_0(u) du = \frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v}.$$

Moments of censoring indicator variable

Under the Weibull baseline survival function, the hazard function for observation  $k$  of individual  $j$  in cluster  $i$  is

$$\lambda_{ijk}(t_{ijk}) = t_{ijk}^\theta \eta_{ij} e^{\beta x_{ijk}} v t_{ijk}^{v-1}.$$

By the assumption that, conditional on the frailties, censoring is not

informative of the frailties (Nielsen et. al., 1992), we have  $\delta_{ijk} | (\xi_i, \eta_{ij}) \sim \text{Poisson}(\mu_{ijk})$ , where

$$\begin{aligned} \mu_{ijk} &= e^{\beta' x_{ijk}} \int_0^{t_{ijk}} u^\theta \eta_{ij} v u^{v-1} du \\ &= e^{\beta' x_{ijk}} \frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v}. \end{aligned}$$

Notice that, for fixed  $j$ ,  $w_{1j}, \dots, w_{nj}$  are independent. Thus  $\text{cov}(w_{il}, w_{jl}) = 0$ , where  $i \neq j$ . For fixed  $i$ ,

$$\begin{aligned} \text{cov}(w_{i1}, w_{i2}) &= \text{cov}(t_{i1k}^\theta \eta_{i1}, t_{i2k}^\theta \eta_{i2}) \\ &= (t_{i1k} t_{i2k}')^\theta \sigma_1 \sigma_2 \rho. \end{aligned}$$

1. Unconditional Mean:

$$\begin{aligned} \tilde{\mu}_{ijk} &= E(\delta_{ijk}) \\ &= E[E(\delta_{ijk} | w_{ij})] = E(\mu_{ijk}) \\ &= e^{\beta' x_{ijk}} E\left(\frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v}\right) \\ &= e^{\beta' x_{ijk}} \frac{v}{\theta + v} t_{ijk}^{\theta + v}. \end{aligned}$$

2. Unconditional Variance:

$$\begin{aligned} \text{var}(\delta_{ijk}) &= E[\text{var}(\delta_{ijk} | w_{ij})] + \text{var}[E(\delta_{ijk} | w_{ij})] \\ &= E(\mu_{ijk}) + \text{var}(\mu_{ijk}) \\ &= e^{\beta' x_{ijk}} \frac{v}{\theta + v} t_{ijk}^{\theta + v} \\ &\quad + e^{2\beta' x_{ijk}} \text{var}\left(\frac{v}{\theta + v} \eta_{ij} t_{ijk}^{\theta + v}\right) \\ &= e^{\beta' x_{ijk}} \frac{v}{\theta + v} t_{ijk}^{\theta + v} + e^{2\beta' x_{ijk}} \left(\frac{v}{\theta + v} t_{ijk}^{\theta + v}\right)^2 \sigma_j^2. \end{aligned}$$



3. Unconditional Correlation (covariance):

If  $k \neq k'$ , note that given  $w_{ij}, T_{ijk}, T_{ijk'}$  are independent and conditional on the frailties, censoring is uninformative of the frailties.

$$\begin{aligned} & \text{COV}(\delta_{ijk}, \delta_{ijk'}) \\ &= E(\text{COV}(\delta_{ijk}, \delta_{ijk'} | w_{ij})) \\ & \quad + \text{COV}[E(\delta_{ijk} | w_{ij}), E(\delta_{ijk'} | w_{ij})] \\ &= \text{COV}(\mu_{ijk}, \mu_{ijk'}) \\ &= e^{\beta'(x_{ijk} + x_{ijk'})} \text{COV}\left(\frac{\nu}{\theta + \nu} \eta_{ij} t_{ijk}^{\theta + \nu}, \frac{\nu}{\theta + \nu} \eta_{ij} t_{ijk'}^{\theta + \nu}\right) \\ &= e^{\beta'(x_{ijk} + x_{ijk'})} \left(\frac{\nu}{\theta + \nu}\right)^2 (t_{ijk} t_{ijk'})^{\theta + \nu} \sigma_j^2. \end{aligned}$$

If  $j \neq j'$ ,

$$\begin{aligned} \text{cov}(\delta_{ijk}, \delta_{ij'k'}) &= 0 + \text{cov}(\mu_{ijk}, \mu_{ij'k'}) \\ &= e^{\beta'(x_{ijk} + x_{ij'k'})} \text{cov}\left(\frac{\nu}{\theta + \nu} \eta_{ij} t_{ijk}^{\theta + \nu}, \frac{\nu}{\theta + \nu} \eta_{ij'} t_{ij'k'}^{\theta + \nu}\right) \\ &= e^{\beta'(x_{ijk} + x_{ij'k'})} \left(\frac{\nu}{\theta + \nu}\right)^2 (t_{ijk} t_{ij'k'})^{\theta + \nu} \sigma_1 \sigma_2 \rho. \end{aligned}$$

If  $i \neq i'$ ,

$$\begin{aligned} & \text{COV}(\delta_{ijk}, \delta_{i'j'k'}) \\ &= \text{COV}(\mu_{ijk}, \mu_{i'j'k'}) \\ &= e^{\beta'(x_{ijk} + x_{i'j'k'})} \text{COV}\left(\frac{\nu}{\theta + \nu} \eta_{ij} t_{ijk}^{\theta + \nu}, \frac{\nu}{\theta + \nu} \eta_{i'j'} t_{i'j'k'}^{\theta + \nu}\right) \\ &= 0 \end{aligned}$$

Thus,  $\delta_{ijk}$ 's can be treated as a sequence of correlated Poisson variables with over-dispersion since the variance of  $\delta_{ijk}$  is not constant.

Appendix 2: Asymptotic properties

As we can see that the variance matrices in equation (1), (4) involve parameters besides  $\beta$ . Consistent estimate of  $\beta$  can be obtained by

replacing  $\theta$  and  $\nu$  with their  $n^{-1/2}$ -consistent estimates (Liang and Zeger, 1986) and the asymptotic properties are well established in this case. As stated in Liang and Zeger (1986), under mild regularity conditions, the estimate of  $\hat{\beta}$  from the generalized estimating equation (1) and (4) is consistent and  $n^{1/2}(\hat{\beta} - \beta)$  is asymptotically multivariate Gaussian as  $n \rightarrow \infty$ , where  $\beta$  is the true value. For the estimates of variance of  $\theta$ ,  $\varphi$  and  $\nu$ , we adopt the grouped jackknife approach because the exact formulae are not available. The estimates are bias corrected and the asymptotic properties for  $\varphi$ , and  $\theta$  will be shown in the following section, thus, we can use Wald type statistic  $\hat{\theta}^2 / \text{var}_J(\hat{\theta})$ , to test the existence of time-modulated frailty, where  $\text{var}_J(\hat{\theta})$  is grouped jackknife variance estimate for  $\hat{\theta}$ .

For the estimate of  $\varphi$  from the penalized score equation (3) or (6), under mild regularity conditions, we have following theorem and give a semi-rigorous proof.

Theorem 1. The estimate  $\hat{\varphi}$  of  $\varphi$  is consistent and  $n^{1/2}(\hat{\varphi} - \varphi)$  is asymptotically normal as  $n \rightarrow \infty$  if  $\max_i(n_i) < M$ , where  $M$  is a known integer.

Proof: Under the true values of  $\beta, \nu$  and  $\varphi$ ,

$$\begin{aligned} & E\left(\frac{U_\varphi}{n}\right) \\ &= \frac{1}{n} E\left\{\sum_{i=1}^n \left(\frac{\partial \tilde{\mu}_i}{\partial \varphi}\right) \text{var}(Y_i)^{-1} (Y_i - \tilde{\mu}_i) + \kappa(n) \sum_{i,k} \frac{\delta_{ik}}{\varphi}\right\} \\ &= \frac{1}{n} \kappa(n) E\left(\sum_{i,k} \frac{\delta_{ik}}{\varphi}\right) = \frac{1}{n} \kappa(n) O(n) = o(1), \end{aligned}$$

as  $n \rightarrow \infty$  since and  $\kappa(n) = n^{-\tau}, \tau > 0$ . By the law of large numbers, we have  $\frac{U_\varphi}{n} - E\left(\frac{U_\varphi}{n}\right) \rightarrow 0$ , in probability as  $n \rightarrow \infty$ . Therefore, from the above two equations,  $\frac{1}{n} U_\varphi = o_p(1)$ . Thus,  $\hat{\varphi}$  is consistent estimate of  $\varphi$ . The asymptotical normality of  $\hat{\varphi}$  can be obtained following the

proof in the appendix of Liang and Zeger (1986). Q.E.D.

Because  $\hat{\nu}$  is moment estimate which is consistent and  $\varphi = \nu + \theta$ , thus,  $\hat{\theta}$  is also consistent.

Appendix 3: Jackknife variance estimation and bias correction

For the parameter  $\beta$ , we can use the robust estimate building in the existing procedure. The parameter  $\theta$  is indicator of the effect of time-modulated frailty, and it is our interest to see whether this effect exist, thus we cannot treat it as a nuisance parameter. First, we notice that the estimate of  $\theta$  is not unbiased because of the penalty term in equation (3) or (6) and

$$E \frac{\partial l}{\partial \theta} \neq 0$$

(Page 28, McCullagh and Nelder, 1983). We will obtain the variance estimate as well an estimation of bias by grouped jackknife method (Therneau and Hamilton, 1997).

The grouped jackknife procedure is the following: Each time we delete the observations from each cluster (or a patient), say cluster  $i$ , and obtain the estimate, say  $\hat{\theta}_{(i)}$ , by applying above estimating procedure to the rest of the data. Let  $\hat{\theta}$  be the estimate based on the all the observations, then the grouped jackknife estimation of variance for  $\theta$  is

$$\text{var}_J(\hat{\theta}) = \frac{(n-1)}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2,$$

$$\text{where } \hat{\theta}_{(.)} = \sum_i \hat{\theta}_{(i)} / n.$$

The bias estimate for  $\theta$  is  $\hat{B}_\theta = (n-1)(\hat{\theta}_{(.)} - \hat{\theta})$ .

Thus, the bias corrected estimate for  $\theta$  is

$$\tilde{\theta} = \hat{\theta} - \hat{B}_\theta = n\hat{\theta} - (n-1)\hat{\theta}_{(.)}.$$

The reason that we apply the grouped jackknife procedure is that we have correlated observations in each cluster and the observations from different clusters are independent.

Theorem 2. Under suitable conditions, the grouped jackknife estimates  $\text{var}_J(\hat{\varphi})$  and  $\text{var}_J(\hat{\theta})$  are asymptotically unbiased estimates of the variance of  $\hat{\varphi}$  and variance of  $\hat{\theta}$ .

Proof: The arguments are similar to Grambsch and Therneau (2000). Q.E.D.

## Accounting For Non-Independent Observations In 2×2 Tables, With Application To Correcting For Family Clustering In Exposure-Risk Relationship Studies

Leslie A. Kalish  
New England Research Institutes  
Watertown, MA

Katherine A. Riester  
Biogen, Inc.  
Cambridge, MA

Stuart J. Pocock  
London School of Hygiene and  
Tropical Medicine  
London, UK

---

Participants in epidemiologic studies may not represent statistically independent observations. We consider modifications to conventional analyses of 2×2 tables, including Fisher's exact test and confidence intervals, to account for correlated observations in this setting. An example is provided, assessing the robustness of conclusions from a published analysis.

Key words: Chi-square test for independence, clustered data, confidence interval, correlation, epidemiologic methods, Fisher's exact test, odds ratio

---

### Introduction

Participants in epidemiologic studies may not represent statistically independent observations. For instance, some individuals may belong to the same family. This will usually make simple statistical tests for exposure-risk relationships anti-conservative, i.e., the strength of evidence for a relationship will be exaggerated by ignoring the lack of independence. We consider a method to modify the standard statistical tests for 2×2 tables in this setting, in order to account for such non-independent observations.

For convenience and clarity, we describe the method in terms of an example comparison of "exposed" and "unexposed" children born to mothers enrolled in a study. Intra-family correlations may induce inter-dependence or clustering of outcomes between siblings. If the exposure of interest is a fixed characteristic of the mother, such as whether or not the mother is

positive for a hereditary gene mutation, then all children of that mother will be concordant on their exposure. This will tend to induce positive correlations between outcomes in the siblings. Other exposures (e.g., gender of the child) may be concordant or discordant, and some exposures (e.g., birth-order) will always be discordant. Most previous research in this area has focused only on settings with no discordant exposures.

In this paper we provide a correction factor for the ordinary Pearson chi square test for independence, and for the construction of confidence intervals, and also propose a method for applying the correction factor to Fisher's exact test. The correction factor depends on the numbers of concordant pairs in each exposure group, the number of discordant pairs, and the intra-family correlation in outcome. We evaluate properties of the new tests using simulations.

An important application of these methods is in evaluating published epidemiologic findings based on a 2×2 table when correlated observations have been naively assumed to be independent. The methods in this paper can then be used to check the robustness of their findings after accounting for non-independence.

---

Leslie A. Kalish is Principal Research Scientist. Email: LesK@neri.org. Katherine A. Riester is Senior Biostatistician. This work was performed while she was employed at New England Research Institutes. Email: Katherine\_Riester@biogen.com. Stuart Pocock is Professor of Medical Statistics. Email: Stuart.Pocock@lshtm.ac.uk.

Methodology

Suppose there are  $N_1$  and  $N_2$  subjects, respectively, in the exposed and unexposed groups ( $N=N_1+N_2$ ). Let  $\hat{\pi}_1$  and  $\hat{\pi}_2$  denote the estimated probabilities of a binary disease outcome in the two groups. Assuming all observations are independent, under the null hypothesis of equal response probabilities,  $H_0: \pi_1 = \pi_2$ , the variance of  $\hat{\pi}_1 - \hat{\pi}_2$  is

$$\text{var}[\hat{\pi}_1 - \hat{\pi}_2] = \pi(1 - \pi)[1/N_1 + 1/N_2], \tag{1}$$

where  $\pi = \pi_1 = \pi_2$ .

Thus the normal approximation statistic for testing  $H_0$  is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\bar{\pi}(1 - \bar{\pi})[1/N_1 + 1/N_2]}}$$

where  $\bar{\pi} = (N_1\hat{\pi}_1 + N_2\hat{\pi}_2)/(N_1 + N_2)$  denotes the overall estimate of response probability from both groups combined.  $Z$  has approximately a standardized normal distribution under  $H_0$  when  $N_1$  and  $N_2$  are large, and  $Z^2$  is the statistic from the ordinary Pearson chi square test for independence.

To account for lack of independence, let  $\rho$  be the within-family correlation of disease outcome (i.e., the correlation between binary variables), which is assumed known. Let  $S$  be the total number of sibling pairs. Note that each individual can be in more than one of the  $S$  pairs, for example four siblings would contribute six pairs to  $S$ . Let  $S_{11}$ ,  $S_{12}$  and  $S_{22}$  denote the number of concordant exposed, discordant and concordant unexposed pairs, respectively (where ‘‘concordant exposed’’ means that both members of the pair are exposed and the other terms are defined similarly). Thus  $S = S_{11} + S_{12} + S_{22}$ . Using standard results for the variance of a linear combination of correlated variables, it can be shown that

$$\text{var}[\hat{\pi}_1 - \hat{\pi}_2] = \pi(1 - \pi) \{1/N_1 + 1/N_2 + 2\rho(S_{11}/N_1^2 - S_{12}/[N_1N_2] + S_{22}/N_2^2)\} \tag{2}$$

Expressions (1) and (2) suggest that the Pearson chi square statistic should be multiplied by a correction factor.

$$CF = \frac{1/N_1 + 1/N_2}{1/N_1 + 1/N_2 + 2\rho(S_{11}/N_1^2 - S_{12}/[N_1N_2] + S_{22}/N_2^2)}$$

We refer to this as the modified chi square test for independence. In practice,  $\rho$  needs to be estimated, or a range of values used, because it is usually unknown.

It seems plausible that the correction factor can also be used to account for non-independence when performing Fisher’s exact test, as would be appropriate in studies with small sample sizes. Suppose one wants an  $\alpha=0.05$  level Fisher’s exact test. Rather than rejecting  $H_0$  when the sum of probabilities of extreme tables is less than 0.05 (which corresponds to rejecting  $H_0$  if Pearson’s chi square statistic is greater than 3.84), one would use the nominal p-value which corresponds to the probability that the chi square distribution exceeds  $3.84 \times CF$ . We refer to this as the modified Fisher’s exact test.

The methods described so far have been in terms of hypothesis testing. By relaxing the null hypothesis assumption that  $\pi_1 = \pi_2$ , one can extend the results so that confidence intervals can be constructed. Generalizing expression (2) by allowing  $\pi_1 \neq \pi_2$  yields the following formula for the variance of the risk difference, which accounts for correlations:

$$\text{var}[\hat{\pi}_1 - \hat{\pi}_2] = \frac{\pi_1(1-\pi_1)}{N_1} + \frac{\pi_2(1-\pi_2)}{N_2} + 2\rho \left[ \begin{array}{c} \frac{\pi_1(1-\pi_1)S_{11}}{N_1^2} \\ - \frac{\sqrt{\pi_1(1-\pi_1)\pi_2(1-\pi_2)}S_{12}}{N_1N_2} \\ + \frac{\pi_2(1-\pi_2)S_{22}}{N_2^2} \end{array} \right].$$

In practice,  $\pi_1$  and  $\pi_2$  would be replaced by observed proportions from the data. Similarly, the familiar variance estimate for the log odds ratio (OR) based on a  $2 \times 2$  table with cell entries  $\{a,b,c,d\}$ , where  $\hat{\pi}_1 = a/(a+b)$  and  $\hat{\pi}_2 = c/(c+d)$ , is

$$\hat{\text{var}}(\log \hat{OR}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

With correlated observations this generalizes to:

$$\hat{\text{var}}(\log \hat{OR}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} + 2\rho \left( \begin{array}{c} \frac{S_{11}}{N_1} \left[ \frac{1}{a} + \frac{1}{b} \right] - \\ \frac{S_{12}}{\sqrt{N_1N_2}} \sqrt{\left[ \frac{1}{a} + \frac{1}{b} \right] \left[ \frac{1}{c} + \frac{1}{d} \right]} + \\ \frac{S_{22}}{N_2} \left[ \frac{1}{c} + \frac{1}{d} \right] \end{array} \right).$$

These variance estimates can be used to construct confidence intervals for a risk difference or odds ratio based on a normal approximation.

We evaluate the true size of the modified Pearson chi square test for independence and modified Fisher's exact tests via simulations. For simplicity, in all simulations we assumed equal

exposure group sample sizes ( $N_1=N_2$ ) and a maximum number of siblings per family of two.

Letting  $\theta$  denote the response probability, consider a fairly rare and a common outcome probability,  $\theta = 0.1$  and  $0.5$ ; small and large intra-family correlations,  $\rho = \{0.2, 0.8\}$ ; three total sample sizes,  $N=N_1+N_2 = \{24, 100, 500\}$ ; and a low and high proportion of  $N$  which is made up of siblings,  $2S/N \approx \{0.08, 0.64\}$ . (A footnote to Table 3 below explains why we were not always able to achieve  $2S/N=0.08$  and  $0.64$  exactly.)

For each combination, we considered three ways that the  $S$  sibling pairs could be divided into concordant exposed, concordant unexposed and discordant pairs, as shown in Table 1. In configurations A and B all sibling pairs are concordant whereas in configuration C all pairs are discordant. Configuration A represents the extreme case where all concordant pairs are in a single exposure group. We did not consider cases with both concordant and discordant pairs because the signs on the  $S_{ij}$  terms in expression (2) show that these terms would tend to cancel each other out and the results would be intermediate between configurations considered.

All combinations of  $\theta$ ,  $\rho$ ,  $N$ ,  $S$  and configurations A-C were simulated (except for combinations with  $\{N=24, \theta=0.1\}$ , which has a substantial probability of a zero marginal total because the study was too small). Thirty thousand simulations for each combination guaranteed that for a true rejection probability of 0.05, we would have a 95% chance of observing a rejection probability within  $[0.0475, 0.0525]$ . In the simulations, we used randomized critical regions (Cox and Hinkley, 1974) to correct for discreteness of the test statistic. Although this may not be used in practice, it makes the different procedures comparable by removing the inherent conservatism in Fisher's exact test (Agresti, 1996).

Example

Dickover et al. (1996) analyzed mother-to-child transmission of human immunodeficiency virus (HIV) in 97 mother-infant pairs, including two pregnancies resulting in twins and three mothers each having two singleton pregnancies. Thus, the 97 mother-infant pairs represented 95 pregnancies in 92 women.

Table 1. Three configurations for allocating siblings to concordant exposed, concordant unexposed and discordant pairs in the simulation study.

		Configuration								
		A			B			C		
		conc	disc		conc	disc		conc	disc	
exp	100	0	100	50	0	50	0	50	50	
unexp	0	0	0	50	0	50	0	50	50	
	100	0	100	100	0	100	0	100	100	

Note: Numbers in each cell represent the percentage of the total number of siblings. (conc = concordant, disc = discordant, exp = exposed, unexp = unexposed).

One of the exposures considered is the use of the antiretroviral treatment zidovudine (ZDV) by the mother during pregnancy and/or during labor and delivery. In all, four of 43 ZDV exposed infants were HIV infected compared with 16 of 54 ZDV unexposed infants. The conventional Pearson  $\chi^2$  statistic without continuity correction is 6.043, corresponding to a two-sided p-value of 0.014. The two-sided Fisher’s exact test p-value is 0.022.

Although we know  $S_{11}+S_{12}+S_{22}=5$ , we have only partial information on the values of  $S_{11}$ ,  $S_{12}$  and  $S_{22}$  from the paper. Clearly, ZDV exposure within each of the twin pairs must be concordant, although we do not know if each pair is exposed or unexposed, leading to the restriction  $S_{11}+S_{22} \geq 2$ . The paper states that ZDV was used in both pregnancies by at least one of the mothers with two singleton births, yielding  $S_{11} \geq 1$ .

Given these restrictions, the most extreme allocations of  $\{S_{11}, S_{12}, S_{22}\}$  result from setting  $\{S_{11}=5, S_{12}=0, S_{22}=0\}$ , or at the other extreme,  $\{S_{11}=1, S_{12}=3, S_{22}=1\}$ . Table 2 shows for both these extremes, the p-values for the modified Pearson  $\chi^2$  and the modified Fisher’s exact test over a range of values for  $\rho$  from  $-1.0$  to  $1.0$ . The  $\rho=0$  column corresponds to the naïve analysis. The true (unknown) correlation is plausibly small and positive, although there are not sufficient data to evaluate this. However, even at the theoretical

extremes ( $\rho=\pm 1.0$ ) the p-values change very little, illustrating that the presence of a small number of correlated observations in this data set has only minimal impact on the statistical findings.

The estimated odds ratio relating HIV infection to ZDV exposure is 0.243 with 95% confidence interval (CI), assuming independence, of (0.075, 0.795). Assuming  $\{S_{11}=5, S_{12}=0, S_{22}=0\}$  and a correlation of  $\rho=.20$ , the CI becomes (0.073, 0.812). With a correlation of  $\rho=1.0$  the CI becomes (0.068, 0.879). Again, the correlation has only minimal impact on statistical findings.

Table 2. Modified Pearson chi square test for independence square p-value (top entry) and modified Fisher’s exact test p-value (bottom values) for mother-to-child HIV transmission example.

$\{S_{11}, S_{12}, S_{22}\}$	$\rho$						
	-1.0	-0.5	-0.2	0.0	0.2	0.5	1.0
{5,0,0}	.008	.011	.013	.014	.015	.017	.021
	.014	.018	.020	.022	.024	.026	.031
{1,3,1}	.015	.014	.014	.014	.014	.013	.013
	.023	.023	.022	.022	.022	.021	.021

Simulation

Simulation results for configurations B and C are shown in Table 3. Because  $N_1=N_2$ , the properties of the different tests are nearly invariant to any allocation of concordant siblings to the exposed and unexposed groups, and hence results for configuration A (not shown) are very similar to configuration B. Both the modified tests perform well, although the modified Fisher’s exact test appears to correct for correlation better than the modified Pearson chi square test in most situations studied.

Table 3. Simulation Results.

N <sup>1</sup>	$\theta^2$	Config <sup>3</sup>	$\rho^4$	2S/N <sup>5</sup>	Actual Test Size for Nominal $\alpha=.05$ Test			
					Pearson	Modified Pearson	Fisher	Modified Fisher
100	.1	B	.2	Low	.0493	.0470	.0496	.0473
100	.1	B	.8	Low	.0598	.0462	.0594	.0520
100	.1	B	.2	High	.0639	.0414	.0635	.0485
100	.1	B	.8	High	.1164	.0519	.1117	.0504
500	.1	B	.2	Low	.0523	.0523	.0537	.0523
500	.1	B	.8	Low	.0557	.0494	.0576	.0503
500	.1	B	.2	High	.0666	.0515	.0673	.0517
500	.1	B	.8	High	.1094	.0498	.1110	.0507
24	.5	B	.2	Low	.0687	.0687	.0541	.0502
24	.5	B	.8	Low	.0804	.0499	.0624	.0489
24	.5	B	.2	High	.0829	.0533	.0647	.0495
24	.5	B	.8	High	.1422	.0729	.1178	.0520
100	.5	B	.2	Low	.0571	.0571	.0500	.0480
100	.5	B	.8	Low	.0650	.0444	.0562	.0492
100	.5	B	.2	High	.0733	.0486	.0650	.0505
100	.5	B	.8	High	.1188	.0481	.1077	.0489
500	.5	B	.2	Low	.0557	.0451	.0508	.0489
500	.5	B	.8	Low	.0624	.0507	.0579	.0499
500	.5	B	.2	High	.0667	.0453	.0614	.0472
500	.5	B	.8	High	.1168	.0541	.1096	.0504

Table 3. (Continued)

N <sup>1</sup>	$\theta^2$	Config <sup>3</sup>	$\rho^4$	2S/N <sup>5</sup>	Actual Test Size for Nominal $\alpha=.05$ Test			
					Pearson	Modified Pearson	Fisher	Modified Fisher
100	.1	C	.2	Low	.0491	.0578	.0489	.0501
100	.1	C	.8	Low	.0440	.0529	.0440	.0517
100	.1	C	.2	High	.0373	.0532	.0377	.0523
100	.1	C	.8	High	.0065	.0635	.0079	.0585
500	.1	C	.2	Low	.0480	.0514	.0499	.0509
500	.1	C	.8	Low	.0403	.0474	.0413	.0487
500	.1	C	.2	High	.0357	.0504	.0367	.0511
500	.1	C	.8	High	.0053	.0502	.0053	.0498
24	.5	C	.2	Low	.0614	.0614	.0487	.0513
24	.5	C	.8	Low	.0577	.0577	.0443	.0535
24	.5	C	.2	High	.0456	.0488	.0352	.0519
24	.5	C	.8	High	.0058	.0497	.0057	.0582
100	.5	C	.2	Low	.0537	.0537	.0471	.0488
100	.5	C	.8	Low	.0489	.0490	.0428	.0501
100	.5	C	.2	High	.0412	.0424	.0364	.0507
100	.5	C	.8	High	.0070	.0614	.0061	.0510
500	.5	C	.2	Low	.0544	.0544	.0501	.0521
500	.5	C	.8	Low	.0508	.0508	.0474	.0549
500	.5	C	.2	High	.0392	.0507	.0361	.0509
500	.5	C	.8	High	.0055	.0465	.0047	.0495

<sup>1</sup>N: Total sample size

<sup>2</sup> $\theta$ : Probability of disease outcome

<sup>3</sup>Config: Configuration of concordant exposed, concordant unexposed and discordant sibling pairs (see Table 1)

<sup>4</sup> $\rho$ : Within-family correlation

<sup>5</sup>2S/N: Number of siblings as a proportion of total sample size. Target low and high values of 2S/N are 0.08 and 0.64. With a small total sample size of N=24, it was not possible to achieve 2S/N=0.08 or 0.64 exactly. For example, in configuration A (Table 1), with one concordant pair, 2S/N=2/24=0.08333 instead of 0.08. Similarly, the actual values of 2S/N for configurations B and C were 0.1667 and 0.0833, respectively. Instead of 0.64, the values of 2S/N were 0.50, 0.6667 and 0.50, respectively, for configurations A, B and C. With N=100 or 500, the only combination where it was impossible to achieve the target values of 2S/N was for {2S/N=0.64, Configuration A}, where allocating 64% of the sample to the exposed group would make N<sub>1</sub> exceed N/2. Thus we used 2S/N=0.50 here.



As expected, the conventional tests tend to be anti-conservative when there are concordant siblings (configurations A and B) and conservative when there are discordant siblings. (configuration C). The conventional Pearson chi square test for independence and Fisher's exact tests perform well when there are <10% siblings in the data set ( $2S/N \approx 0.08$ ), even with correlation as high as 0.8. The magnitude of conservatism or anti-conservatism increases with the correlation ( $\rho$ ) and as the number of sibling pairs in the data set ( $S$ ) increases.

### Conclusion

We have presented modifications to the ordinary Pearson  $\chi^2$  test for independence and to Fisher's Exact test for the analysis of  $2 \times 2$  tables when some of the observations are correlated. The methods achieve the desired properties across a wide range of possible data sets even with quite small sample sizes. Formulae for constructing modified confidence intervals are also provided.

Previous work has focused on unstratified and stratified  $2 \times 2$  tables and clustered data where it is assumed that exposure status is common to all units in a cluster (i.e., no discordant pairs) (Donald & Donner, 1987; Donner, 1989; Rao & Scott, 1992; Rosner, 1982). This would occur, for example, if the exposure of interest was a genetic characteristic of the mother of children in a cluster. This assumption is not required in other research (Rosner & Milton, 1988; Begg, 1999) but these methods require enough clustered observations to allow the nature of the correlation to be estimated from the data.

Another possible approach to analysis would be to use a logistic regression model with correlation between siblings from the same family. Standard errors that take the correlation into account can be obtained using generalized estimating equations (Diggle, Liang & Zeger, 1994). Advantages of this modeling approach are that additional covariates can be added to the model, the covariates can be specific to each cluster unit and the exposures of interest need not be dichotomous. However, its complexity is a problem and since it requires availability of the raw data it could not ordinarily be used to evaluate published results.

Our modified procedures require knowledge of the correlation,  $\rho$ , which would be difficult to estimate unless the number of pairs is large. However, by repeating the analysis over a range of possible values for  $\rho$ , one can assess the sensitivity of conclusions to the presence of correlation.

Determining a reasonable range of plausible values for  $\rho$  is difficult in part because correlations of binary variables have unusual properties. It is known that the correlation between binary variables is constrained by the true probabilities as follows (Prentice, 1988): If  $\pi_1 < \pi_2$  then

$$\max \left[ -\sqrt{\frac{\pi_1 \pi_2}{(1-\pi_1)(1-\pi_2)}}, -\sqrt{\frac{(1-\pi_1)(1-\pi_2)}{\pi_1 \pi_2}} \right] \\ < \rho < \sqrt{\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}} .$$

Estimates of  $\pi_1$  and  $\pi_2$  can therefore aid in setting bounds on  $\rho$ . Published results from analyses that naively assumed independence can easily be checked in such a sensitivity analysis, provided one is given enough information about the numbers of pairs in which both pair members are exposed, both are unexposed and exposure status is discordant. Unlike the value of  $\rho$ , these numbers would ordinarily be known when analyzing one's own data but may not be known when assessing the impact of non-independence on published results, in which case a range of possible numbers can be used in a sensitivity analysis.

Although the methods here are presented as for epidemiologic risk relations, they could also apply to clinical trials in which some (but not necessarily all) subjects have more than one "outcome," for example on two eyes in ophthalmologic studies, two legs in studies of walking impairment or multiple teeth in dental studies.

The procedures in this paper are most useful when there is a small amount of clustering so that the correlation cannot be reliably estimated, and when it is desired to evaluate the robustness of

conclusions to deviations from the assumption of independence.

In conclusion, it is important to recognize that non-independent observations, such as subjects within the same family, may make conventional statistical analyses based on independence assumptions prone to be conservative or anti-conservative. Simple correction methods, such as that described here for dichotomous exposure and outcome, are of value in ensuring that appropriately valid inferences are drawn when non-independent observations are present.

#### References

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley and Sons.
- Begg M. D. (1999). Analyzing k (2 x 2) tables under cluster sampling. *Biometrics*, 55, 302-307.
- Cox D. R., Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.
- Dickover, R. E., Garratty, E. M., Herman, S. A., & et al. (1996). Identification of levels of maternal HIV-1 RNA associated with risk of perinatal transmission. Effect of maternal zidovudine treatment on viral load. *Journal of the American Medical Association*, 275, 599-605.
- Diggle, P. J., Liang, K-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. New York: Oxford University Press.
- Donald, A., & Donner, A. (1987). Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine*, 6, 491-499.
- Donner, A. (1989). Statistical methods in ophthalmology: An adjusted chi-square approach. *Biometrics*, 45, 605-611.
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44, 1033-1048.
- Rao J. N. K., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*, 48, 577-585.
- Rosner, B., & Milton, R. C. (1988). Significance testing for correlated binary outcome data. *Biometrics*, 44, 505-512.
- Rosner, B. (1997). Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. *Biometrics*, 38, 105-114.
- Zhang, J., & Boos, D. D. (1997). Mantel-Haenszel test statistics for correlated binary data. *Biometrics*, 53, 1185-1198.

## The Statistical Modeling Of The Fertility Of Chinese Women

Dudley L. Poston, Jr.  
Department of Sociology  
Texas A&M University

---

This article is concerned with the statistical modeling of children ever born (CEB) fertility data. It is shown that in a low fertility population, such as China, the use of linear regression approaches to model CEB is statistically inappropriate because the distribution of the CEB variable is often heavily skewed with a long right tail. For five sub-groups of Chinese women, their fertility is modeled using Poisson, negative binomial, and ordinary least squares (OLS) regression models. It is shown that in almost all instances there would have been major errors of statistical inference had the interpretations of the results been based only on the results of the linear regression models.

Key words: Poisson, negative binomial, OLS, modeling Chinese fertility

---

### Introduction

The national censuses of many countries include a question that asks women about the number of children they have had ever born to them; these are referred to as children ever born (CEB) data. Demographers often use such data in statistical models of fertility. CEB data may be referred to as event count or count data. "An event count refers to the number of times an event occurs... (and) is the realization of a nonnegative integer-valued random variable" (Cameron & Trivedi, 1998, p. 1). For many count variables, such as the CEB variable, its distribution is heavily skewed with a long right tail. This is certainly the case in low-fertility populations, such as China, the population analyzed in this article. This reflects the fact that most women in such populations have children at the lower parities, including zero parity, and few

have children at the higher parities. In this paper CEB data from the 1990 census of China are analyzed for five sub-groups of ever-married women. It is shown that the use of linear regression to model CEB for these sub-groups is statistically inappropriate.

Table 1 (all tables and figures are in the appendix) is a compilation of descriptive information on the CEB variable for ever-married Chinese women aged 15-49 from five sub-groups, namely, the Han (the majority nationality group), and four of China's 55 minority groups (the Korean, Manchu, Hui and Uygur minorities). The Han women have an average of 2.13 children ever born. The Korean and Manchu women have mean CEB values that are less than that of the Han, both with values of 1.8. Hui women have a mean CEB of 2.33, and Uygur women report one of the higher average CEB values of any of the Chinese minority nationalities, a mean of 3.16. Tables and figures appear at the end of this paper.

Figures 1-5 (appendix) show frequency distributions of the observed CEB data (the blue lines with circles as symbols) for these five sub-groups: Han women (Figure 1), Manchu women (Figure 2), Korean women (Figure 3), Uygur women (Figure 4), and Hui women (Figure 5). For Han women (Figure 1), about 8 percent have no children, over 30 percent have one, about 29 percent have two, 19 percent have three, 9 percent have four, 4 percent have five, and progressively

---

Dudley L. Poston, Jr. is Professor of Sociology, and the George T. and Gladys H. Abell Professor of Liberal Arts, at Texas A&M University. He has co-authored/edited ten books and over 200 refereed journal articles, chapters and reports on various sociological, statistical and demographic topics. He is currently co-editing (with Michael Micklin) the *Handbook of Population*, scheduled to be published in 2004 by Kluwer Academic/Plenum. Email: dudleyposton@yahoo.com.

smaller percentages of women have children at the higher parities. The Han distribution is heavily skewed with a long right tail. This characterization also applies to the Manchu, Korean and Hui distributions. Only the Uygur women (Figure 4), with one of the highest fertility rates in China, do not show as skewed a CEB distribution as the others, although their distribution too has a long right tail.

A major point is that none of the distributions in Figures 1-5 is normally distributed, and most are heavily skewed, and all have long right tails. Therefore, the statistical modeling of these kinds of CEB data should be based on approaches other than the ordinary least squares (OLS) linear regression model. Using an OLS model to predict a count outcome, such as children ever born, will often “result in inefficient, inconsistent, and biased estimates” (Long, 1997, p. 217) of the regression parameters.

#### Methodology

There are several alternative models that take into account the characteristics of a count variable such as CEB. The most basic is the Poisson regression model in which “the probability of a count (of CEB) is determined by a Poisson distribution, where the mean of the distribution is a function of the independent variables” (Long, 1997, pp. 217-218), which, in this case would be the characteristics of the individual women. The Poisson regression model, and alternate models such as the negative binomial regression model and some types of zero-inflated regression models, are based on the univariate Poisson distribution, which will now be considered.

#### The Univariate Poisson Distribution

Figures 1-5 also show for the five sub-groups of Chinese women the univariate Poisson distributions (the purple lines with triangle symbols) that correspond to the mean CEB values for the respective groups. The shape of the univariate Poisson distribution depends entirely on the value of the mean, and is based on the following formula:

$$\Pr(Y = y) = \frac{\exp(-\mu)\mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

where the parameter  $\mu$  represents the mean, and

$y$  is an integer indicating the number of times the count has occurred, ranging from 0 to some higher positive integer.

This purely theoretical distribution was developed by the French mathematician Simeon-Denis Poisson (1781-1840) and is fundamental in the statistical analysis of an assortment of issues involving radioactivity, traffic, and many other count events that occur in time and/or space.

Some properties of the theoretical Poisson distribution are (Long & Freese, 2001, p. 224):

1) With increasing values of the mean,  $\mu$ , the shape of the distribution moves to the right; this is seen in the above CEB distributions;

2) The variance of the univariate Poisson distribution equals the mean,  $\mu$ , a property known as equi-dispersion. Empirically, however, the variance of many count variables tends to be greater than the mean. To illustrate, the descriptive CEB data in Table 1 indicate that the variance of CEB for Uygur women is more than twice its mean. The variance of CEB for Hui women is also larger than its mean.

3) As  $\mu$  increases, the probability of zero counts decreases.

4) As  $\mu$  increases, the Poisson distribution approximates a normal (Gaussian) distribution.

Consider once again Figures 1-5. Observe their empirical distributions of children ever born, and compare these distributions with the univariate Poisson distributions that correspond to their mean CEB values. For Han women (Figure 1), the fitted Poisson distribution (the purple line with triangle symbols) slightly over-predicts the observed proportion of women with zero children, under-predicts the proportion with one child, slightly under-predicts the proportion with two children, and predicts fairly well the proportions of women at the higher parities. The univariate Poisson distributions for the other four nationality groups of Chinese women also show various patterns of under-prediction and over-prediction of the numbers of women at most of the different counts of children ever born. In some cases these patterns of under- and over-prediction are similar to those

of the Han Chinese shown in Figure 1, and in other cases they are not.

One should not expect the univariate Poisson distributions to perfectly predict the proportions of women at each count of CEB because the Poisson distributions do not take into account the heterogeneity of the women. That is, one reason why the Poisson distributions shown in Figures 1-5 do not perfectly fit the observed CEB distributions is that the women in the five samples vary in the numbers of children they produce. It would be unrealistic to expect that all Han women have the same rate of child production, that all Manchu women have the same rate, and similarly for the other groups of women. The researcher needs to introduce heterogeneity into the models by drawing on the observed characteristics of the women. Therefore, the issue of statistical modeling will now be considered and the results of the analyses presented.

### Results

Most demographic analyses of CEB have used linear regression models (e.g., see Ritchey, 1975; Johnson, 1979; Janssen and Hauser, 1981; Entwisle and Mason, 1985; Bean and Tienda, 1987). This is an appropriate statistical strategy if the mean CEB count is high because in such a situation the distribution of the dependent variable tends to be approximately normal. But if the mean of the counts is not high, as is the case with children ever born responses of women in low-fertility populations, then the "common regression estimators and models, such as ordinary least squares in the linear regression model, ignore the restricted support for the dependent variable" (Cameron & Trivedi, 1998, p. 2).

There is a host of regression models that may be used in the analysis of count data (see Cameron & Trivedi, 1998). The Poisson regression model is the most basic and the standard model for analyzing count outcomes and is derived from the Poisson distribution. The Poisson regression model is an appropriate strategy when the mean and the variance of the count distribution are similar, and is less applicable when the variance of the distribution exceeds the mean, that is, when there is over-dispersion in the count data. In such instances an

alternate modeling approach would be negative binomial regression.

### The Poisson Regression Model

In a Poisson regression model, the dependent variable, namely, the number of events, i.e., the number of children ever born, is a nonnegative integer and has a Poisson distribution with a conditional mean that depends on the characteristics (the independent variables) of the women. The model thus incorporates observed heterogeneity according to the following structural equation:

$$\mu_i = \exp(a + X_{1i} b_1 + X_{2i} b_2 + \dots + X_{ki} b_k)$$

where:  $\mu_i$  is the expected number of children ever born for the  $i^{th}$  woman;  $X_{1i}$ ,  $X_{2i}$  ...  $X_{ki}$  are her characteristics; and  $a$ ,  $b_1$ ,  $b_2$  ...  $b_k$  are the Poisson regression coefficients.

The Poisson regression model is a nonlinear model, predicting for each individual woman the number of children she has had ever born to her,  $\mu_i$ . The  $X$  variables are related to  $\mu$  nonlinearly. Some applications of the Poisson regression model will now be illustrated in separate statistical analyses of children ever born for Han, Korean, and Manchu women, using data from the 1% Sample of the 1990 Census of China. The Chinese samples have been restricted to ever-married women between the ages of 15 to 49. Poisson models would appear to be appropriate for estimating CEB for the Han, Manchu and Korean because their mean and variance CEB values are so similar (Table 1).

A selection of independent variables is used that reflect socioeconomic and locational characteristics that have been shown to be associated with fertility. The independent variables pertain to age, education, residence, regional location, and marital status. Some are measured as dummy variables, and others as interval. They are the following:  $X_1$  is the woman's age measured in years (age);  $X_2$  to  $X_5$  are four dummy variables representing the levels of education of the women, namely,  $X_2$ , completed at least some elementary school;  $X_3$ , completed at least some middle school;  $X_4$ , completed at least some high school; and  $X_5$ , completed at least some college; illiterate women are treated as the reference group;  $X_6$  is the

woman's employment status, a dummy variable coded 1 if she is employed;  $X_7$  and  $X_8$  are dummy variables representing the woman's residence in a city (yes/no) and her residence in a town (yes/no); the reference category is residing in a rural area;  $X_9$  to  $X_{13}$  are five dummy variables representing the woman's region of residence, namely,  $X_9$  residence in the North,  $X_{10}$  residence in the East,  $X_{11}$  residence in the South Central,  $X_{12}$  residence in the Southwest, and  $X_{13}$  residence in the Northwest; residence in the Northeast region is treated as the reference category; and  $X_{14}$  and  $X_{15}$  are two dummy variables reflecting the woman's marital status as follows:  $X_{14}$  indicates if the woman is widowed (yes or no), and  $X_{15}$  if she is divorced (yes or no); currently married is the reference category.

The Poisson regression model is estimated with maximum likelihood procedures. Table 2 reports the results of the above Poisson regression model for Han women, Manchu women and Korean women. All three models converged after three iterations. The overall structure of the models may be appraised with the Likelihood Ratio  $\chi^2$  statistic, which tests the null hypothesis ( $H_0$ ) that all the Poisson coefficients are not significantly different from zero. In all three models the null hypothesis may be rejected, indicating that there is some predictive utility in the three models. This conclusion is reinforced by the significant values of the three Pseudo  $R^2$  statistics.

The decision to use a Poisson regression approach to model CEB for the Han, Manchu and Korean women may be formally and directly appraised with the Poisson Goodness of Fit  $\chi^2$  test statistic (bottom of Table 2); it compares the observed empirical distribution with the distribution predicted by the Poisson regression model. The null hypothesis ( $H_0$ ) is that there is no difference between the observed data and the modeled data, indicating that the Poisson model fits the data. A small  $\chi^2$  value, with a probability  $> 0.05$ , indicates that one cannot reject the null hypothesis that the observed CEB data are Poisson distributed. In all three models, the values of the Poisson Goodness of Fit  $\chi^2$  statistic indicate that using Poisson regression to model the CEB data was appropriate.

The Poisson regression coefficients for the fifteen independent variables will now be

examined. Table 2 reports for each independent variable the value of the Poisson coefficient ( $b$ ) and its standard error (*s.e.*). Coefficients that are not significant have been asterisked. The Poisson coefficients indicate the degree of nonlinear association of the independent variable with the dependent variable of CEB, controlling for the effects of the other independent variables.

Looking first at the model for Han women, age is positively associated with CEB. And the four education dummy variables are negatively associated with CEB (the reference variable here is illiterate status). If the woman is employed ( $X_6$ ), she has fewer children than if she is not employed. Women who live in cities ( $X_7$ ), or in towns ( $X_8$ ), have fewer children than women who live in rural areas. Women who live in the North ( $X_9$ ), or in the South Central ( $X_{11}$ ), or in the Southwest ( $X_{12}$ ), or in the Northwest ( $X_{13}$ ) have more children than women living in the Northeast region (the reference region). The CEB of women living in the East ( $X_{10}$ ) is not significantly different from the CEB of women living in the Northeast. The CEB of widowed women ( $X_{14}$ ) is not significantly different from the CEB of married women, but the CEB of divorced women ( $X_{15}$ ) is significantly less than that of married women. None of the signs of the Poisson coefficients are surprising. They are what one would expect.

The effects of the Poisson coefficients for the independent variables in the other two regression models, those for Manchu women and for Korean women, are quite similar in sign, and in magnitude as those for Han women. However, more of the coefficients in the Manchu and Korean models are not statistically significant compared to the number of insignificant coefficients in the Han model. Five of the fifteen coefficients in the Manchu regression model are not significant (four of the region variables, and the widowed variable). And eleven of the fifteen coefficients in the Korean model are not statistically significant; only the age, college, city residence, and divorced variables are statistically significant.

It was noted earlier in the review of the demographic literature on the statistical modeling of children ever born that many CEB analyses have used linear regression approaches. It was also noted that such a strategy is not appropriate in low fertility populations owing to the heavily skewed

distribution of CEB. One thus might ask how similar, or different, would the regression results reported in Table 2 be if linear regression models had been used instead of Poisson regression models.

Table 3 reports ordinary least squares regression results for the same Han, Manchu and Korean populations using the same independent and dependent variables. There are many differences between the OLS regression results shown in Table 3 and the Poisson regression results shown in Table 2. The most important differences have to do with the statistical significance of many of the coefficients. For instance, in the equations for the Han women, and in the equations for the Korean women, more OLS coefficients are statistically significant than are the corresponding Poisson coefficients. In the two Manchu equations, the same five coefficients do not achieve statistical significance.

Among Han women all the OLS coefficients are significant, whereas two of their corresponding Poisson coefficients are not significant. Among the Korean women, six of their fifteen OLS coefficients are not significant, but eleven of their Poisson coefficients are not significant.

Had an OLS model, instead of a Poisson model, been used to predict the number of children ever born among Korean women, incorrect statistical inferences would have been made for the effects of five of the fifteen variables. The results of the OLS model would have allowed the inferences that Korean women who have completed middle school ( $X_4$ ), and high school ( $X_5$ ), have fewer children than Korean women who are illiterate. In the Poisson regression these coefficients are not significant. Also, the OLS regression results permit the inferences that employed Korean women ( $X_5$ ) have fewer children ever born than unemployed Korean women, and women living in towns ( $X_8$ ) have a lower CEB than women living in rural areas; these are two more erroneous statistical inferences. And, according to the OLS results, it would have been concluded that women living in the South Central region ( $X_{11}$ ) have more children ever born than women living in the Northeast region, another incorrect inference.

Poisson regression models were estimated for Han, Manchu and Korean women because their

mean and variance values for CEB were similar (Table 1). However, Poisson regression models were not estimated for the Hui and Uygur women because their respective variance CEB values were larger than their corresponding mean CEB values (Table 1) indicating the apparent presence for each group of over-dispersion in their CEB distributions.

If there is significant over-dispersion in the distribution of the count (CEB) variable for a population, the estimates from the Poisson regression model will be consistent, but inefficient. "Further the standard errors from the (Poisson regression model) will be biased downward, resulting in spuriously large z-values" (Long, 1997, p. 230), which could lead the investigator to make incorrect statistical inferences about the significance of the variables. This situation is addressed by extending the Poisson regression model by adding "a parameter that allows the conditional variance of (the count outcome) to exceed the conditional mean" (Long, 1997: 230). This extension of the Poisson regression model is the negative binomial regression model, which is now considered.

#### The Negative Binomial Regression Model

It was noted earlier that the Poisson regression model "accounts for observed heterogeneity (i.e., observed differences among sample members) by specifying the (predicted count,  $\mu$ ) as a function of the observed" independent variables (Long & Freese, 2001, p. 243). Often, however, the Poisson regression model does not fit the observed data because of over-dispersion. "That is, the model underestimates the amount of dispersion in the outcome" (Long & Freese, 2001, p. 243). In the negative binomial regression model, variation in  $\mu$  "is due both to variation in (the independent variables) among the individuals (in the sample population) and to unobserved heterogeneity introduced by  $\epsilon$ " (Long, 1997, p. 231). The term  $\epsilon$  is a "random error that is assumed to be uncorrelated with (the independent variables) ... ( $\epsilon$  may be thought of) "either as the combined effects of unobserved variables that have been omitted from the model or as another source of pure randomness" (Long, 1997, p. 231).

The negative binomial regression model thus adds to the Poisson regression model the error

term  $\varepsilon$  according to the following structural equation:

$$\mu_i = \exp(a + X_{1i} b_1 + X_{2i} b_2 + \dots + X_{ki} b_k + \varepsilon_i)$$

It may be shown that the distribution of the observations in the negative binomial regression model is still Poisson. In the negative binomial regression model, the mean structure is the same as in the Poisson regression model, but the distribution about the mean is not the same (Long, 1997, p. 233; Long & Freese, 2001, p. 243). If there is not a statistically significant amount of dispersion in the count outcome data, then the negative binomial regression model will reduce to the Poisson regression model.

One way, therefore, to test for dispersion in the count outcome is to estimate a negative binomial regression model along with a Poisson regression model, and to compare the results of the two models. Like the Poisson regression model, the negative binomial regression model is estimated by maximum likelihood procedures.

As already noted, given a data-set with over-dispersion, if one were to estimate both Poisson and negative binomial regression models, both will have the same mean structure. But the Poisson model will tend to under-estimate the dispersion in the dependent variable. Hence, "the standard errors in the Poisson regression model will be biased downward, resulting in spuriously large z-values and spuriously small p-values" (Long & Freese, 2001, p. 243; Cameron & Trivedi, 1986, p. 31). Also, in the negative binomial model, compared to the Poisson regression model, there will be an increased probability of both low and high counts.

The left panel of Table 4 contains the results of a negative binomial regression model using the fifteen independent variables to estimate the number of children ever born for ever-married Hui women. For comparison purposes, the middle panel of the table contains the results of a Poisson regression estimating Hui CEB using the same independent variables. And in the right panel are presented the results from an OLS regression.

Comparing the values of the negative binomial regression coefficients (left panel of Table 4) with the values of the Poisson regression coefficients (middle panel), it may be seen that the

two sets of coefficients are virtually identical. This suggests that there is not a significant amount of dispersion in the CEB data for the Hui women.

The formal statistical test for appraising the presence of dispersion in the negative binomial distribution is the parameter, alpha (in the Poisson regression model, thus, alpha = 0). (See StataCorp, 2001, volume 2, p. 386-387, 390-391; Long & Freese, 2001, p. 243-245 for more discussion.) At the bottom of Table 4 (left panel) is the value of alpha, and immediately below it, the likelihood-ratio  $\chi^2$  test of alpha. The value of alpha is .000, indicating that there is not a statistically significant amount of dispersion in the distribution of CEB for the Hui women. The likelihood ratio  $\chi^2$  test of alpha has a value of .000, with a probability of .5.

This  $\chi^2$  test is based on a comparison of the value of the final log likelihood from the negative binomial regression model and the corresponding value from the Poisson model. There is no difference in the values, indicating that the CEB data for the Hui women are Poisson distributed. This conclusion is reinforced by the results of the Poisson Goodness of Fit of Fit  $\chi^2$  (bottom of the middle panel of the table), which has a probability of 1.0. This means that the Poisson model fits the data; the Poisson goodness of fit  $\chi^2$  test indicates that given the Poisson regression model one cannot reject the null hypothesis that the observed data are Poisson distributed.

Before leaving the CEB regressions for the Hui women, the Poisson results will be compared with the OLS regression results. What kinds of inference errors would have been made had the Hui CEB been estimated with a linear regression model? The results of the OLS regression model would have allowed the conclusion that among the Hui women employment status ( $X_6$ ) has a significant negative effect on CEB. Thus it would have been inferred that employed women have fewer children ever born than women who are not employed. This turns out to be an incorrect inference. The Poisson regression model results indicate no statistical relationship between employment status and CEB.

Similar errors of inference would have been regarding the effects on CEB of the woman's location in the East region ( $X_{10}$ ), the South Central region ( $X_{11}$ ), and the Northwest region ( $X_{13}$ ). For all three of these regional location variables the



OLS regression results indicate that the effects are significant, but the Poisson regression results show they are not. The Poisson regression model is the more statistically appropriate approach for modeling CEB among the Hui women.

Finally, the estimation of children ever born among the Uygur women may be considered. For Uygur women the variance of their CEB is more than twice the magnitude of the mean of their CEB, values of 6.99 and 3.16, respectively. Table 5 presents in the left panel the results of a negative binomial regression model estimating Uygur CEB, along with the results of a Poisson regression model in the center panel, and the results of an OLS regression model in the right panel. The first question is whether there is enough over-dispersion in Uygur CEB to justify the use of a negative binomial regression model.

The first indication that the negative binomial model is appropriate is the fact that the coefficients from the model are very different from the corresponding coefficients from the Poisson model. A second and more formal indication is that alpha, the over-dispersion parameter (bottom of the table, left panel), has a value of .113, with a probability of .005. And the likelihood-ratio  $\chi^2$  test of alpha has a high value of 776.0, with a probability of .000, indicating that the probability that one would observe these data if the process was Poisson, i.e., if  $\alpha = 0$ , is virtually zero. The Uygur data are clearly not Poisson. A final and related indication is that the Poisson Goodness of Fit  $\chi^2$  test statistic performed on a Poisson regression of the Uygur CEB data (bottom of the middle panel of the table) has a probability of .000. This means that the Poisson model does not fit the data; according to the Poisson goodness of fit  $\chi^2$  test, the null hypothesis that the observed data are Poisson distributed must be rejected.

The negative binomial and Poisson coefficients (Table 5) may now be compared. First, the signs of the effects of the independent variables on CEB are all the same. Also, the six predictors that are not statistically significant in one model are not significant in the other model. However, for thirteen of the independent variables, the standard errors in the Poisson model are smaller than those in the negative binomial model (the standard errors for the age variable ( $X_1$ ) are the same in both models). This means that for

thirteen of the fourteen independent variables, in the Poisson model the z-values will be spuriously high and the p-values spuriously low. Although there would have been no errors of statistical inference had these Poisson regression results, rather than the negative binomial regression results, been used to predict Uygur CEB, the potential for error is much greater using the Poisson results. For all the above reasons, the negative binomial model is the preferred regression model for predicting children ever born among Uygur women.

Finally, the results of the negative binomial regression predicting Uygur CEB may be compared with the OLS results (left and right panels of Table 5). Would any inference errors be committed had the OLS results been used? The major error that would have occurred is with regard to the effect on CEB of employment status. The results of the OLS regression model indicate that among Uygur women employment status ( $X_6$ ) has a statistically significant negative effect on CEB. Thus one would have inferred that employed Uygur women have fewer children ever born than Uygur women who are not employed, controlling for the effects of the other independent variables. This turns out to be an incorrect inference. The negative binomial regression results show no statistical relationship between employment status and CEB. Some of the implications of the research reported in this paper will now be addressed.

### Conclusion

This article considered distributions of CEB data for five sub-groups of Chinese women. It was shown that they were not normal (Gaussian), but, rather, heavily skewed with long right tails. Such distributions are characteristic of low-fertility populations. Given such distributions, a linear regression model is inappropriate for the statistical modeling of children ever born. Fifteen socioeconomic and locational variables drawn from the 1990 Census of China were then used as independent variables to model CEB for the Han and minority group women.

For the Han and Manchu and Korean women, both Poisson regression and ordinary least squares (OLS) regression models were estimated. And for the Hui and Uygur women, these same two approaches along with negative binomial

regression were used. It was shown that in almost all instances there would have been major errors of statistical inference had the interpretations been based only on the results of linear regression models.

The literature on the statistical modeling of CEB data indicates that in many instances, linear regression models have been used. The decision to use a linear model, however, is only appropriate if the average CEB value is high. When the mean of a count outcome is high, say, at least above 4 or 5, but certainly around 8 or 9, then the distribution of the outcome will often tend to be approximately normal. However, few populations these days, except mainly those in sub-Saharan Africa, have fertility this high. It would appear thus that the use of a linear model for modeling a fertility variable such as children ever born is becoming more and more inappropriate. And in low fertility populations, such as China, using a linear model would clearly be inappropriate statistically.

References

Bean, F. D., Tienda, M. (1987). *The Hispanic population of the United States*. New York: Russell Sage Foundation.

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1, 29-53.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, U.K.: Cambridge University Press.

Entwisle, B., & Mason, W. M. (1985). Multilevel effects of socioeconomic development and family planning programs on children ever born. *American Journal of Sociology*, 91, 616-649.

Janssen, S. G., & Hauser, R. M. (1981). Religion, socialization, and fertility. *Demography*, 18, 511-528.

Johnson, N. E. (1979). Minority-group status and the fertility of Black Americans, 1970: A new look. *American Journal of Sociology*, 84, 1386-1400.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, California: Sage Publications.

Long, J. S., & Freese, J. (2001). *Regression models for categorical dependent variables using stata*. College Station, Texas: Stata Press.

Ritchey, P. N. (1975). The effect of minority group status on fertility: A re-examination of concepts. *Population Studies*, 29, 249-257.

StataCorp. (2001). *Stata statistical software: release 7.0*. College Station, Texas: Stata Corporation.

Appendix: Tables and Figures

Table 1: Descriptive Data for Children Ever Born: Ever-Married Han, Manchu, Korean, Uyгур, and Hui Women, Ages 15-49, China, 1990

Group	Standard			No. of Cases
	Mean	Dev.	Variance	
Han	2.1326	1.4202	2.0170	216,312
Manchu	1.8047	1.1745	1.3795	20,210
Korean	1.7959	1.0478	1.0978	3,837
Uyгур	3.1577	2.6443	6.9921	14,553
Hui	2.3289	1.7662	3.1194	17,976

Source of Data: 1% Sample of the 1990 Census of China. The sample of Han women is a 1/10 sample of the 1% sample.

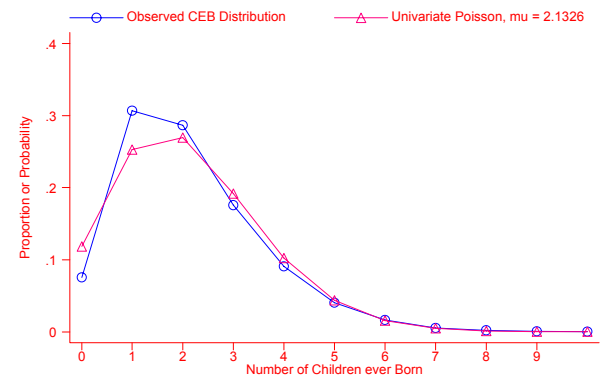


Fig. 1: CEB Dist. for the Han and Poisson Dist. with mu = 2.1326

Table 2: Poisson Regression Models Predicting Number of Children Ever-born for Ever-Married Han, Manchu and Korean Women, Ages 15-49, China, 1990

	Han	Manchu	Korean
Sample Size	216,312	20,210	3,837

Independent Variable	Han		Manchu		Korean	
	b	s.e.	b	s.e.	b	s.e.
X <sub>1</sub> Age	.050	.000	.055	.001	.052	.002
X <sub>2</sub> Elem. Sch	-.092	.004	-.076	.019	.005	.085*
X <sub>3</sub> Middle Sch	-.239	.005	-.189	.020	-.058	.085*
X <sub>4</sub> High School	-.353	.007	-.248	.025	-.117	.089*
X <sub>5</sub> College	-.583	.020	-.466	.054	-.301	.123
X <sub>6</sub> Employ Status	-.063	.005	-.095	.012	-.013	.031*
X <sub>7</sub> City Residence	-.398	.006	-.335	.022	-.234	.038
X <sub>8</sub> Town Residence	-.096	.004	-.055	.013	-.029	.028*
X <sub>9</sub> North Region	.018	.007	.050	.013	-.099	.086*
X <sub>10</sub> East Region	-.003	.006*	-.055	.069*	-.045	.181*
X <sub>11</sub> S. Central Reg.	.120	.006	.014	.054*	.152	.134*
X <sub>12</sub> SW Region	.034	.007	.060	.097*	-.182	.290*
X <sub>13</sub> NW Region	.089	.008	-.029	.071*	-.032	.236*
X <sub>14</sub> Widowed	-.022	.012*	-.040	.048*	-.023	.071*
X <sub>15</sub> Divorced	-.285	.028	-.261	.081	-.341	.129
Constant	-.809	.010	-1.057	.034	-1.145	.111
Pseudo R <sup>2</sup>	.145	.000	.136	.000	.112	.000
Likelihood Ratio $\chi^2$	106740.4	0.00	8456.5	0.00	1283.5	0.00
Poisson Goodness of Fit $\chi^2$	106486.4	1.00	7527.8	1.00	1322.9	1.00

\*Coefficient not significant at p < .05.

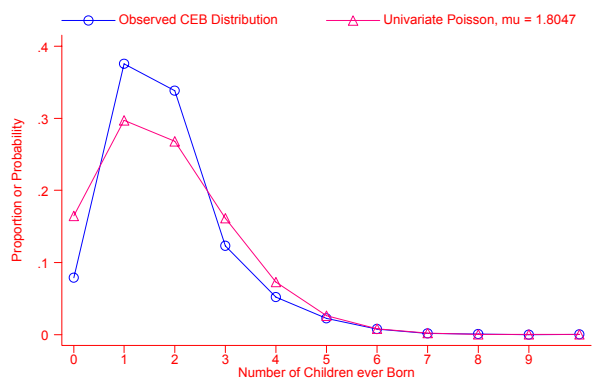


Fig. 2: CEB Dist. for the Manchu and Poisson Dist. with  $\mu = 1.8047$

Table 3: Ordinary Least Squares Regression Models Predicting Number of Children Ever-born for Ever-Married Han, Manchu and Korean Women, Ages 15-49, China, 1990

	Han	Manchu	Korean
Sample Size	216,312	20,210	3,837

Independent Variable	Han		Manchu		Korean	
	b	s.e.	b	s.e.	b	s.e.
X <sub>1</sub> Age	.110	.000	.106	.001	.095	.002
X <sub>2</sub> Elem Sch	-.311	.006	-.277	.023	-.028	.097*
X <sub>3</sub> Middle Sch	-.569	.007	-.478	.024	-.220	.096
X <sub>4</sub> High Sch	-.727	.009	-.591	.027	-.333	.098
X <sub>5</sub> College	-.975	.021	-.858	.048	-.521	.117
X <sub>6</sub> Employ Stat	-.192	.007	-.224	.012	-.062	.030
X <sub>7</sub> City Resid	-.762	.007	-.557	.020	-.383	.034
X <sub>8</sub> Town Resid	-.223	.006	-.110	.013	-.067	.027
X <sub>9</sub> North Region	.023	.009	.091	.013	-.144	.077*
X <sub>10</sub> East Region	-.019	.008	-.119	.063*	-.100	.175*
X <sub>11</sub> S. Cent Reg	.262	.008	.025	.052*	.293	.143
X <sub>12</sub> SW Region	.082	.009	.106	100*	-.370	.234*
X <sub>13</sub> NW Region	.189	.011	.013	.067*	-.001	.233*
X <sub>14</sub> Widowed	.096	.021	.074	.062*	-.060	.081*
X <sub>15</sub> Divorced	-.482	.032	-.354	.067	-.492	.095
Constant	-.951	.014	-1.012	.036	-1.066	.115
R <sup>2</sup> (adj.)	.531	.000	.577	.000	.559	.000
F-test	16293.0	.000	1839.1	.000	1283.5	.000

\*Coefficient not significant at p < .05.

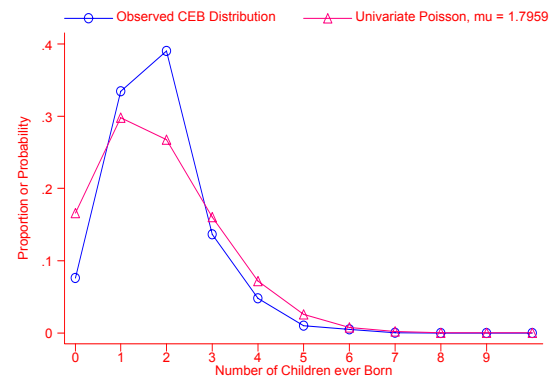


Fig. 3: CEB Dist. for the Koreans and Poisson Dist. with  $\mu = 1.79$

Table 4: Negative Binomial Regression Model (NBR), Poisson Regression Model (PR), and Ordinary Least Squares Regression Model (OLS) Predicting Number of Children Ever-born for 17,976 Ever-Married Hui Women, Ages 15-49, China, 1990

Independent Variable	NBR Model		PR Model		OLS Model		
	b	s.e.	b	s.e.	b	s.e.	
X <sub>1</sub> Age	.054	.001	.054	.001	.133	.001	
X <sub>2</sub> Elem Sch	-.108	.014	-.108	.014	-.412	.026	
X <sub>3</sub> Middle Sch	-.234	.017	-.234	.017	-.559	.029	
X <sub>4</sub> High Sch	-.341	.024	-.341	.024	-.674	.037	
X <sub>5</sub> College	-.575	.062	-.575	.062	-.998	.079	
X <sub>6</sub> Employ	-.013	.017*	-.013	.017*	-.081	.031	
X <sub>7</sub> City Res.	-.354	.017	-.354	.017	-.828	.027	
X <sub>8</sub> Town Res.	-.072	.017	-.072	.017	-.225	.029	
X <sub>9</sub> North Reg.	-.024	.026*	-.024	.026*	-.061	.040*	
X <sub>10</sub> East Reg.	.047	.029*	.047	.029*	.117	.045	
X <sub>11</sub> S. Cent Reg.	.048	.029*	.048	.029*	.109	.045	
X <sub>12</sub> SW Reg.	-.008	.030*	-.008	.030*	-.039	.049*	
X <sub>13</sub> NW Reg.	.287	.026	.286	.026	.689	.042	
X <sub>14</sub> Widowed	-.029	.037*	-.029	.037*	.084	.083*	
X <sub>15</sub> Divorced	-.490	.058	-.490	.058	-.913	.081	
Constant		-.959		.039		-1.742	.066
Pseudo R <sup>2</sup> / R <sup>2</sup> (adj.)		.181		.000		.189	.000
Likelihood Ratio $\chi^2$ or							
F-test		12072.2		.000		12763.3	.000
Alpha		.000		.000			
L-Ratio $\chi^2$ test of alpha				.000		.500	
Poisson Goodness of Fit $\chi^2$				11049.0		1.000	

\*Coefficient not significant at p < .05.

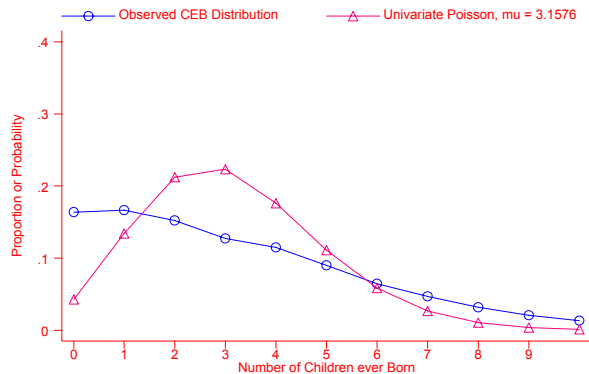


Fig. 4: CEB Dist. for the Uygur & Poisson Dist. with mu = 3.1576

Table 5: Negative Binomial Regression Model(NBR), Poisson Regression Model (PR), and Ordinary Least Squares Regression Model (OLS) Predicting Number of Children Ever-born for 14,553 Ever-Married Uygur Women, Ages 15-49, China, 1990

Independent Variable	NBR Model		PR Model		OLS Model		
	b	s.e.	b	s.e.	b	s.e.	
X <sub>1</sub> Age	.060	.001	.057	.001	.184	.002	
X <sub>2</sub> Elem Sch	.059	.014	.071	.012	.213	.045	
X <sub>3</sub> Middle Sch	.071	.018	.085	.015	.202	.055	
X <sub>4</sub> High Sch	-.074	.026	-.060	.022	-.196	.075	
X <sub>5</sub> College	-.259	.070	-.234	.061	-.608	.183	
X <sub>6</sub> Employ	-.019	.016*	-.025	.013*	-.121	.048	
X <sub>7</sub> City Res.	-.247	.021	-.248	.018	-.817	.061	
X <sub>8</sub> Town Res.	-.052	.019	-.060	.016	-.232	.056	
X <sub>9</sub> N Region	-.076	.049*	-.103	.067*	.289	2.326*	
X <sub>10</sub> E Region	.147	.899*	.117	.817*	.798	2.253*	
X <sub>11</sub> S. Cent Reg.	.218	.830*	.195	.750*	1.091	2.113*	
X <sub>12</sub> SW Region						variable not included	
X <sub>13</sub> NW Region	.649	.783*	.629	.707*	2.116	2.014*	
X <sub>14</sub> Widowed	-.202	.035	-.183	.028	-.608	.117	
X <sub>15</sub> Divorced	-.800	.032	-.802	.029	-1.426	.066	
Constant		-1.480		.784*		-1.348	.708*
Pseudo R <sup>2</sup> / R <sup>2</sup> (adj.)		.190		.000		.421	.000
Likelihood Ratio $\chi^2$ or							
F-test		8042.6		.000		13645.7	.000
Alpha		.113		.005			
L-Ratio $\chi^2$ test of alpha				776.0		.000	
Poisson Goodness of Fit $\chi^2$				21413.4		.000	

\*Coefficient not significant at p < .05.

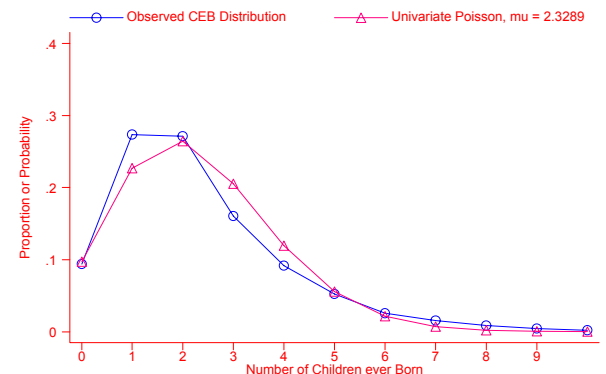


Fig. 5: CEB Dist. for the Hui & Poisson Dist. with mu = 2.3289

## Simulation Study Of Chemical Inhibition Modeling

Pali Sen  
Department of Mathematics and Statistics  
University of North Florida

Mary Anderson  
Department of Mathematics and Statistics  
University of North Florida

---

The combined effects of the activities of different chemicals are of interest of this study. We simulate for the synthetic data, and fit experimental data for three models and estimate the parameters. We assess the fit of the synthetic data and the experimental data by comparing the coefficients of variation for the parameter estimates and identify the best model for the inhibition process.

Key words: Additive model, coefficient of variation, combination model, product model

---

### Introduction

Pharmacological data deal with the study of chemicals in a body. Researchers are interested in the distributions of these chemicals and their retention times. Studies by clinicians (e.g., Wagner, 1988; Bass, 1988; Beck, 1988) on the specific activities of chemicals under various conditions are examples. Thakur (1988), Matis (1988), and Jacquez (1985), to name a few, developed methods to study the dynamic behavior of chemicals using tools in mathematical modeling.

Sen and Mohr (1990), and Sen, Bell, and Mohr (1992) studied the distribution of a chemical in a body and modeled its activities as nonlinear time-dependent functions. In this paper we develop mathematical models of two chemicals in order to study the inhibition effects of one chemical on the other. This inhibition between two chemicals may be indicated by suppression or amplification of their individual effects. The specific activities of two interacting chemicals are

measured on laboratory animals during an experiment.

Three models are developed here for study: an additive model, a product model, and a combination model. The purpose of the study is to select the best model from these three models, to describe the inhibition effect of two interacting chemicals and to interpret the observed data. A simulation study of the models and their parameter estimation using the synthetic data is described in the result section. A numerical example of the evaluation of the models is also presented in the result section.

### Methodology

Consider a chemical flow in a body and its concentration changes at different times and at different points. We observe the flow discretely at a certain location in the body and at certain times, and we visualize a one-compartment model with a single input and output from the system. After the initial dose of a chemical is injected into the system, some amount of it will escape the compartment and the chemical itself will slowly decay over time. We assume the rate changes in concentration,  $p(t)$ , of the chemical at any time in the body will follow the differential equation given below.

$$dp(t)/dt = -\alpha p(t) + f(t), \quad (2.1)$$

where  $\alpha$  is the rate at which the absorbed chemical leaves the system.  $f(t)$  is a decreasing function of

---

Contact information for both authors is: Department of Mathematics and Statistics, University of North Florida Jacksonville, FL, 32224. Telephone: (904) 620-2846, Fax No. (904) 620-2818. E-mail: [psen@unf.edu](mailto:psen@unf.edu). The authors thank the referees for many good suggestions on content and presentation.

the chemical applied initially, which enters the system and is assumed to have the form

$$f(t) = d e^{-\beta t}, \quad (2.2)$$

where  $d$  is the initial amount of the input, and  $\beta$  is the rate of absorption of the chemical. The solution of the equation (2.1) may be extended for two chemicals, since they follow essentially the same equation. Hence the solution of equation (2.1) for each chemical is written as,

$$p_i(t) = d_i (\exp(-\beta_i t) - \exp(-\alpha_i t)) / (\alpha_i - \beta_i), \quad (2.3)$$

for  $i = 1, 2$ .

We now consider an 'activator-inhibitor' system for the combined concentrations,  $p(t)$ , of the activity levels, which consists of two chemicals that each exhibits the mutual effect of inhibiting the other's formation, Edelstein - Keshet, (1989). By selecting models for each of the combining effects, we have models that take the following forms:

$$\text{Model 1: } p(t) = p_1(t) - p_2(t). \quad (2.4)$$

$$\text{Model 2: } p(t) = p_1(t) * p_2(t). \quad (2.5)$$

$$\text{Model 3: } p(t) = p_1(t) - p_2(t) + p_1(t) * p_2(t) \quad (2.6)$$

The rationale for these models is based on the physiological combination effects of two chemicals. Sometimes the combined effects produce a reduction, and at other times a surge in the activity levels, depending on the chemical balance of the concentration levels. The negative sign in (2.4) indicates inhibition of the first chemical by the second, which is an antagonistic effect. Next, we consider the product model since the combination may alternatively cause the effects to rise. The product of the two equations is similar to an interaction effect, which we believe is a competitor for model 1. The third model is a combination of models 1 and 2, which intuitively may be viewed as a synergistic effect. We want to achieve a trend to identify a best inhibition model using experimental and synthetic data.

Computationally, the proposed models in (2.4), (2.5), and (2.6) yield different combinations of exponential terms. To simplify the notations, we use  $\alpha, \beta, \gamma, \delta$  instead of  $\alpha_1, \beta_1, \alpha_2, \beta_2$ . Here,  $\gamma$  represents the rate at which the second chemical leaves the system and  $\delta$  is the rate at which the second chemical is absorbed in the system. The initial input ( $d_i$ ) is considered to be of the same amount,  $d$ , for both the chemicals. We write equations (2.4), (2.5), and (2.6) in the following equations.

$$p(t) = d[\exp(-\beta t) - \exp(-\alpha t)] / (\alpha - \beta) - d[\exp(-\delta t) - \exp(-\gamma t)] / (\gamma - \delta). \quad (2.7)$$

$$p(t) = d^2[\exp(-(\beta t + \gamma t)) - \exp(-(\beta t + \delta t)) - \exp(-(\alpha t + \gamma t)) + \exp(-(\alpha t + \delta t))] / (\alpha - \beta)(\delta - \gamma). \quad (2.8)$$

$$p(t) = d[\exp(-\beta t) - \exp(-\alpha t)] / (\alpha - \beta) - d[\exp(-\delta t) - \exp(-\gamma t)] / (\gamma - \delta) + d^2[\exp(-(\beta t + \gamma t)) - \exp(-(\beta t + \delta t)) - \exp(-(\alpha t + \gamma t)) + \exp(-(\alpha t + \delta t))] / (\alpha - \beta)(\delta - \gamma). \quad (2.9)$$

The above equations are similar even though the combinations of the parameters are different in each equation. Each equation in (2.7) – (2.9) consists of four parameters. We compare the fit of the generated curves with the observed values and then study the errors of estimation for each fitted curve.

## Results

We want to compare the models by generating data from the respective equations for a period of time. We simulate the models with four unknown parameters and for thirteen time points.  $d$  is a proportionality constant and may be set to any number. A value of  $d = 10$  units is considered for the analysis. The random numbers are generated for ten sets of data at each time point 0, 30, ...360. The system of random numbers is perturbed by a sigma of 1 unit. The Monte Carlo method of the program is written using Fortran language and the Levenberg -Marquardt is used to fit the model parameters (Press, 1986). The initial guesses of the parameters and the first derivatives of the parameters are supplied in order for the nonlinear equations to converge when a chi-square value has

reached to a pre set number. Convergence implies that the best estimates of the parameters have been obtained, under the assumption that the model is adequate. Two convergence criteria are used here.

- 1) Continue iterative method until the parameter values on successive iterations stabilize. This can be measured by the size of the each parameter increment relative to the previous parameter value.
- 2) Continue till relative change in sum of squares on successive iterations is small.

Compliance with both criteria does not guarantee convergence; instead it could indicate a lack of progress. Often a small pivot element will generate a large correction in the parameter values, which will then be rejected. This near degeneracy of the minimum causes the parameters to fluctuate around a value (a local minimum) without ever converging to a global minimum.

Table 1 gives the estimated parameter values along with their standard errors for the data generated using the additive model for initial estimates of the parameters  $\alpha = .0699$ ,  $\beta = .0173$ ,  $\gamma = .3742$ , and  $\delta = .057$ , with respective parameter estimates  $\hat{\alpha} = .0958$ ,  $\hat{\beta} = .00535$ ,  $\hat{\gamma} = .420862$ ,  $\hat{\delta} = .0228$ . The change in the Chi-Squares is from 186326.5 to 110324.7 with a 41% drop in the value.

Table 1 -Parameter estimates for three models for the first set of simulated data± indicates asymptotic standard errors

Model	$\alpha$	$\beta$	$\gamma$	$\delta$
<b>Additive</b>	.096± .000028	.005± .000001	.421 ± .00018	.023 ± .000056
<b>Product</b>	.0019± 26.5040	.0083± 26.5040	15.19± 26.5040	-.0014± 26.5040
<b>Combination</b>	5.816± .000516	.00009± .000004	.0002± .0000015	.0078± .000019

Table 2 gives the estimated parameter values along with their standard errors for the data generated using the combination model for initial estimates of the parameters  $\alpha = .0818$ ,  $\beta = .0108$ ,

$\gamma = .0114$ , and  $\delta = .114$  with respective parameter estimates  $\hat{\alpha} = .845261$ ,  $\hat{\beta} = .00622$ ,  $\hat{\gamma} = .00669$ ,  $\hat{\delta} = 3.145268$ . The change in the Chi-Squares is a 99% drop in the value.

Table 2 -Parameter estimates for three models for the second set of simulated data± indicates asymptotic standard errors.

Model	$\alpha$	$\beta$	$\gamma$	$\delta$
<b>Additive</b>	.1396± .00012	.0004± .00003	.3087 ± .00043	.0015 ± .00003
<b>Product</b>	.0016± 77.223	.3669± 77.229	5.445± 78.636	-.0013± 77.224
<b>Combination</b>	.845± .000939	.006± .00003	-.007± .00003	3.145± .00503

Tables 1 and 2 show some similarity in the estimates of the parameters. We have obtained the convergence criteria by all three models for the above two sets of parameters. It was extremely difficult to find the initial estimates of the parameters for the product model, but we included it in the analysis as well. The additive and the combination models both gave very good estimates of the standard errors, but the product model had the estimated standard errors very large to indicate the convergence might have reached locally. The data were generated using the additive and the combination models and both sets of data converged for both models 1 and 3 with good sets of parameter estimates, but neither set worked well for the product model. The coefficients of variation for estimated parameters fitted from the simulation data were calculated by dividing the standard errors of estimation by the estimated parameters for the sets given in the accompanying tables.

Once the validity of the models has been established, we want to see how the three models compare at each other, we use the estimated parameter values from the tables to draw the curves for all three models and place them on the same axes. Figure 1 shows that all three graphs basically follow the same pattern but in figure 2 the product model shows a slight fluctuation from the other two curves, and the combination model

separates from the other two at the end of 360 minutes. These pictures confirm that all three models are equally good in describing the chemical inhibition process.

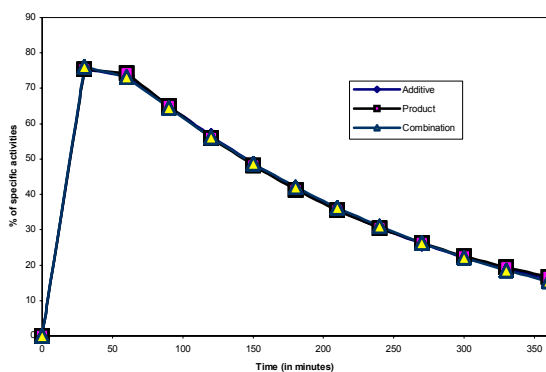


Figure 1. Simulated curves for three models using the parameter estimates in Table 1.

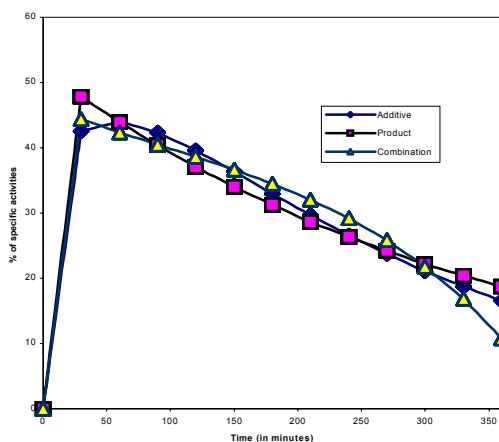


Figure 2. Simulated curves for three models using the parameter estimates in Table 2.

The simulation study is convincing enough for us to look further into the models using the real data. The data used for this study were collected at the Ohio State University pharmacological laboratory in Columbus, Ohio. Researchers administered two chemicals, morphine and midazolam, to laboratory rats. The experiment is to study the effects of two chemicals, Midazolam and Morphine when they are administered simultaneously. A high dose of Morphine, a common anesthetic agent, may have

an irreversible side effect on the body. Midazolam has been shown to either increase or decrease spinal activity depending on the relative combined concentration of morphine and midazolam, Niv (1988); Tejwani, (1990). Also midazolam has been shown to have minimal side effects even with high dosages.

The purpose of their study for the combination effects was to see the effects of morphine in high doses when applied with varying dose levels of midazolam. Researchers especially want to determine if a combination level of two chemicals can produce the desired anesthetic effect that reaches high within 50 minutes to 100 minutes and gets out of the system within 3 hours. The experimenters used a group of five to six laboratory rats to administer midazolam at three levels and morphine at the same three levels as a 3X3 factorial design.

The combined effects of those two chemicals were observed on the rats. The concentration levels for each chemical were used at 10 $\mu$ g (low), 20 $\mu$ g (moderate), and 30 $\mu$ g (high) and each of the nine combinations of the concentrations. The numbing effects of the combined chemicals were recorded by measuring the tail flickering of the rats. These measurements, known as the specific activities, represent the percentage increase over the baseline values of the anesthetic effects, which are due to the chemicals. Higher measurement readings indicate a stronger effect of the chemicals.

The average percentages of the maximal possible effects on tail flickering of these animals were measured. A high number indicated the effect of analgesia (anesthetic effect) was strongly present. A descriptive study of the data has been published in one of the pharmacological journals, Rattan (1991).

Nonlinear regression fits of the models to the data are obtained using the Marquardt method. The estimates of the parameters are also obtained. The procedure is iterative based on the least squares method. The initial guess for each parameter is supplied and a known value of the initial amount (d) of 10 units is used for each level of the chemicals for the observed thirteen time points. The coefficients of variation for estimated parameters fitted from the data are calculated for the converged sets.



To avoid repetition and lack of any further meaningful information, only three selected combination levels of midazolam and morphine are presented here. The tables 3, 4, and 5 show the estimates of the four parameters with their corresponding asymptotic standard errors of estimation.

A well-known result is that the method of maximum likelihood asymptotically produces an estimated density, which is closest to the true density in the information sense. Maximizing the log-likelihood is equivalent to minimizing the expected logarithmic difference between the two densities. Akaike (1974) has suggested an estimate of the approximate loss between the true normal density and the approximating density. This estimate uses the maximum log-likelihood of the observation vector minus the number of parameters. Akaike's information criterion (AIC) is a useful statistic for statistical model evaluation and has been widely accepted in some areas of statistics, Bozdogan (1987). It is calculated for each selected model as  $AIC = (n)\ln(SSE^s/n) + 2k$ , SAS (1990). A low value for AIC indicates a better fit.

We notice in table 5, the combination data of both high levels of concentrations (Mor30 and Mid30), fit with AIC values equal to 28.89 for the additive model, and 34.07 for the combination model, those are the smallest among all other AIC values. The AIC values are in the similar range in the table 3 for the combination data of low morphine with high midazolam concentrations (Mor10 and Mid30). For the combination data of medium morphine with low midazolam concentrations (Mor20 and Mid10) in table 4, the AIC values are relatively high but similar for the additive model and the combination model and even higher for the product model.

We compare the standard errors of the parameter estimates in these tables. In tables 3 and 4 only the combination model has reliable estimated standard errors, and in table 5 models 1 and 3 have reliable estimated standard errors. So the combination model is the only one that is holding steady for the data.

Table 3 -Parameter estimates of three models for low level of Morphine $\pm$  indicates asymptotic standard errors. \* = Concentration Level.

Level*	$\alpha$	$\beta$	$\gamma$	$\delta$	AIC
<b>Mor10</b>	.0383 $\pm$	.0382 $\pm$	.3771 $\pm$	.3765 $\pm$	56.42318072
<b>Mid30</b>	2.469	2.4645	617.19	616.3	
<b>Model 1</b>					
<b>Mor10</b>	.2005 $\pm$	.1748 $\pm$	.0001 $\pm$	-.1400 $\pm$	53.15877737
<b>Mid30</b>	0.0000	263.9	27.961	74.52	
<b>Model 2</b>					
<b>Mor10</b>	.0809 $\pm$	.0168 $\pm$	.0120 $\pm$	.1431 $\pm$	53.46542876
<b>Mid30</b>	.0423	.0178	.0152	.0676	
<b>Model 3</b>					

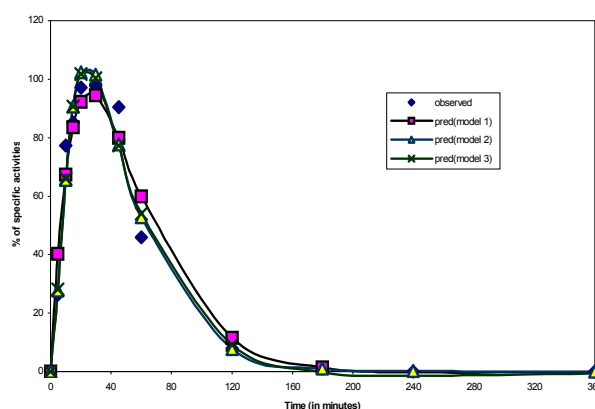


Figure 3. Distribution of Morphine 10 $\mu$ g and Midazolam 30 $\mu$ g with predicted models.

Table 4 -Parameter estimates of three models for medium level of Morphine  $\pm$  indicates asymptotic standard errors.

Level*	$\alpha$	$\beta$	$\gamma$	$\delta$	AIC
<b>Mor20</b>	.0836 $\pm$	.0027 $\pm$	67739 $\pm$	47398 $\pm$	64.50291081
<b>Mid10</b>	.0050	.0005	.0000	.0000	
<b>Model 1</b>					
<b>Mor20</b>	-.0286 $\pm$	.4917 $\pm$	.0299 $\pm$	.4951 $\pm$	74.11086129
<b>Mid10</b>	.0016	6.972	0.0000	7.8581	
<b>Model 2</b>					
<b>Mor20</b>	.1445 $\pm$	.0012 $\pm$	.0094 $\pm$	.1828 $\pm$	63.81358576
<b>Mid10</b>	.0876	.0008	0.0130	.1063	
<b>Model 3</b>					

Note: \* = Concentration Level.

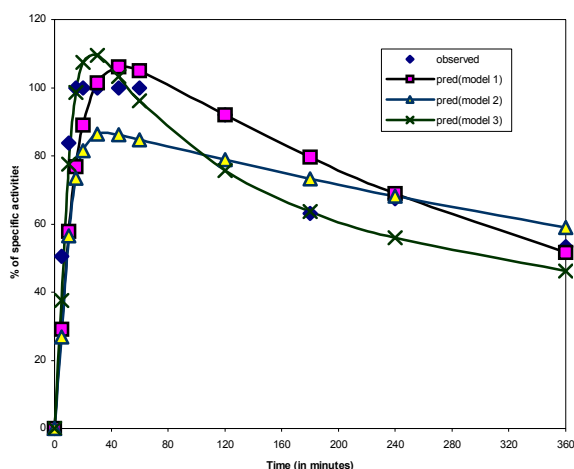


Figure 4. Distribution of Morphine 20µg and Midazolam 10µg with predicted models.

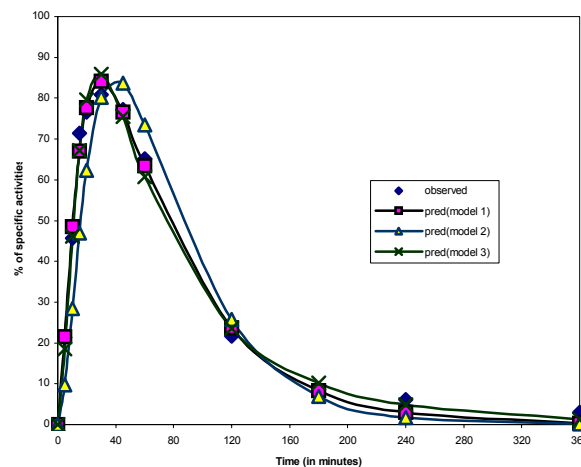


Figure 5. Distribution of Morphine 30µg and Midazolam 30µg with predicted models.

Table 5 -Parameter estimates of three models for high level of Morphine± indicates asymptotic standard errors.

Level*	$\alpha$	$\beta$	$\gamma$	$\delta$	AIC
Mor30 Mid30 Model 1	.0699 ± .0082	.0173 ± .0020	.3742 ± .1141	.0570 ± .0489	28.89092708
Mor30 Mid30 Model 2	.0796 ± 0.0000	.0705 ± 14528	.0288 ± 701.31	-.0446 ± 1396	68.71982797
Mor 30 Mid 30 Model 3	.0818 ± .0286	.0108 ± .0235	.0114 ± .0396	.1141 ± .0267	34.07428277

Note: \* = Concentration Level.

Figures 3 – 5, refer to the respective tables 3 - 5, show the actual data with the estimated fitted lines by the models 1, 2, and 3. The estimated parameter values from the tables are used to draw the respective fitted curves and placed them with the original data points. Figure 3 shows a very close fit by all three curves, figure 4 shows very different fit by all three of them and figure 5 again shows very good fit by all three models.

We now focus on the estimated values to decide how good these fits are. Tables 3 - 5 show a lack of reliability in the measurements of the coefficients of variation by the product model for all of its estimated parameter values. They are quite large, indicating that the convergence may have reached locally, which is also the case with the simulation results for the product model, even though it fit the experimental data in figures 3 and 5. Table 3 shows only the combination model with a set of reasonable coefficients of variation for its estimated parameter values but all curves fit data well. The standard errors for estimated parameter values for the other two models are large in Table 3. For the combination and addition models in table 5, the parameter estimates are extremely good with mostly low coefficients of variation, and all three models fit well. The estimated standard errors with the low coefficients of variation may be used to make the confidence intervals for the parameters for the combination model.

### Conclusion

The AIC criteria has been criticized in literature for adding two times the number of parameters of the model in the calculation, but we overcome this criticism by having equal number of parameters for each model. The AIC values are used heavily in the literature for model comparisons, but how low is a value to be considered for a good fit. Our studies show that the values range from 28.89 to 74.11 for the set of data that we have used. It is then reasonable to suggest that this range of AIC values meet the standards since they meet the convergence criteria for the study.

However to select a best model, only the AIC criteria may not be enough, the estimated parameter values also play a key role in determining a good model. One does not need to do the testing of hypothesis to decide if the estimated values are acceptable or not, as the coefficients of variation are instant indicators for the decision. The coefficients of variation for the estimated parameters are always large for the product model, but they are low for the combination model with no exception, indicating that the combination model is probably a better choice. This indicates that the coefficients of variation should also be considered for the choice of a model.

When we look into the simulation of the models, we find that all three models generate extremely similar patterns. The data under study contain a lot of variations for measurements and has only thirteen time points for each set. This may contribute to some of the convergence problems for model 1, which sometimes produces unusable estimates of the parameters in tables 3 and 4. Otherwise the simulation results in tables 1 and 2 are perfectly fine for the additive model. The combination model always did extremely well for fitting the data, estimating the parameters with low coefficients of variations, but producing the AIC values similar to the other two models.

This study indicates that there are a number of conceivable reasons why a particular model should be chosen. Beyond the reasonable AIC values, we looked into the fit and the coefficients of variation for estimating the parameters. This study showed that the reliable estimates of the parameter values were obtained from the combination model always, from the

additive model sometimes and none of the times from the product model. The fit of the models are extremely close in two of the three graphs shown here. The models 1 and 2 have the potential for simpler interpretation of an inhibition model as being either an additive or a multiplicative in nature, but as we have seen the estimated parameter values are not always reliable, whereas a combination of the two models produces reliable estimates of the parameters.

In conclusion we would like to remark that AIC criteria are a very simple technique to identify the goodness of fit, but we need other statistical techniques as well to evaluate a model. This paper addresses the issue to identify a model that will best describe the inhibition process, even though that may not be a flawless model for the entire process. The models are based on simple approach to the physical description of the inhibition process with a few parameters. The data we have used for the numerical example may be modeled by much complicated equations than these models can describe. Any chemical interaction is a complicated process but the observable data points are restricted. Moreover, this type of experiment requires live subjects for study, which makes it harder to collect a large set of data. The proposed models have only four parameters to estimate and require a moderate size of the data set. In real experimental process if more data is available, the initial equation set up must be more elaborate before the three proposed models could be introduced. The simulation results and the numerical example show that the combination model better describe the inhibition effects of two chemicals.

### References

- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723.
- Bass, L. (1988). Saturable drug uptake by the liver: Models, experiments and methodology. In (A. Pecile & A. Rescigno, Eds.): *Pharmacokinetics, mathematical and statistical approaches to metabolism and distribution of chemicals and drugs*, A., 291-322. Plenum Press.

- Beck, J. S. (1988). Conceptual foundations and uses of models in pharmacokinetics. In (A. Pecile & A. Rescigno, Eds.): *Pharmacokinetics, mathematical and statistical approaches to metabolism and distribution of chemicals and drugs, A*, 11-18. Plenum Press.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370.
- Edelstein-Keshet, L. (1989). *Mathematical models in biology*. New York: Random House.
- Jacquez, J. A. (1985). *Compartmental analysis in biology and medicine*. Ann Arbor: University of Michigan Press, 54 (8), 594-604.
- Matis, J. H. (1988). An introduction to stochastic compartmental models in pharmacokinetics. In (A. Pecile & A. Rescigno, Eds.): *Pharmacokinetics, mathematical and statistical approaches to metabolism and distribution of chemicals and drugs, A.*, 113-128. Plenum Press.
- Niv, D., Davidovich S., Geller E. & Urca G. (1988). Analgesic and hyperalgesic effects of Midazolam: Dependence on route of administration. *Anesth. Analg.*, 67, 1169 - 1173.
- Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986). *Numerical recipes*. New York: Cambridge University Press.
- SAS/STAT User's guide (1990). Ver. 6(2) Cary, North Carolina: SAS Institute Inc.
- Sen, P., & Mohr, D. (1990). A kinetic model for calcium distribution. *Journal of Theoretical Biology*, 142, 179 - 188.
- Sen, P., Bell, D., & Mohr, D. (1992). A calcium model with random absorption: A stochastic approach. *Journal of Theoretical Biology*, 154, 485 - 493.
- Thakur, A.K. (1988). Modeling of pharmacokinetic data. In (A. Pecile & A. Rescigno, Eds.): *Pharmacokinetics, mathematical and statistical approaches to metabolism and distribution of chemicals and drugs, A.*, 27-60. Plenum Press.
- Tejwani, G. A., Rattan, A. K., & McDonald, J. S. (1990). Effect of intrathecal injection of midazolam on morphine induced antinociception in the rat. In (J. M. VanRee, A. H. Mulder, V. M. Wiegant, & T. B. VanWimersa Greidanus, Eds.) *Excerpta medica*. New York , 29 - 31.
- Rattan, A. K., McDonald, J. S., & Tejwani G. A. (1991). Differential effects of intrathecal midazolam on morphine-induced antinociception in the rat: Role of spinal opioid receptors," *Anesth. Analg.*, 73, 124 - 131.
- Wagner, J. G. (1988). Pharmacokinetic Studies in man. In (A. Pecile & A. Rescigno, Eds.): *Pharmacokinetics, mathematical and statistical approaches to metabolism and distribution of chemicals and drugs, A.*, 291 - 322.

## Combining Quantum Mechanical Calculations And A $\chi^2$ Fit In A Potential Energy Function For The $\text{CO}_2 + \text{O}^+$ Reaction

Ellen F. Sawilowsky  
Detroit, Michigan

---

In order to compute a highly accurate statistical rate constant for the  $\text{CO}_2 + \text{O}^+$  reaction, it is necessary to first calculate the potential energy of the system at many different geometric configurations. Quantum mechanical calculations are very time-consuming, making it difficult to obtain a sufficient number to allow for accurate interpolation. The number of quantum mechanical calculations required can be significantly reduced by using known relations in classical physics to calculate energy for configurations where the oxygen is relatively far from the  $\text{CO}_2$ . A chi-squared fit to quantum mechanical points is obtained for these configurations, and the resulting parameters are used to generate an equation for the potential energy. This equation, combined with an interpolated set of quantum mechanical points to give the potential energy for configurations where the molecules are closer together, allows all configurations to be calculated accurately and efficiently.

Key words: Potential energy surface,  $\chi^2$  fit

---

### Introduction

The reaction of carbon dioxide with the  $\text{O}^+$  oxygen ion is of interest because experimental rate measurements show that at low energies the rate is constant at the expected value, but at high energies the rate steadily decreases to values below the expected rate (Viggiano, et al., 1992). RRKM rate calculations were done for the purpose of explaining this experimentally observed decrease (Forst, 1973).

In order to calculate the rate of reaction using statistical rate theories such as RRKM theory, the potential energy of the reacting molecules must be known at any geometric configuration that might be found near the transition state. This refers to the small portion of the potential surface that is near the maximum point on the minimum-energy path.

The accuracy of a rate calculation is directly related to the accuracy of the potential surface employed, and a good potential is needed if the rate calculation is to be highly accurate. Because calculating the potential energy at any one configuration involves time-consuming quantum mechanical calculations, constructing the potential surface with energies for all probable configurations near the transition state using quantum mechanical calculations becomes an impossible task. Instead, it is common to do calculations at judiciously chosen configurations and use interpolation to obtain good approximations for the energies of configurations for all other geometries.

The potential is split into long and short-range portions in order to further reduce the number of quantum mechanical calculations. *Ab initio* quantum mechanical calculations were done for the short-range portion only. At separation distances of 6.9 Å or greater, the long-range portion of the potential is invoked. It consists of a fit to the long range *ab initio* points with a functional form, which is a parameterized variation of the ion-induced dipole plus quadrupole potential:

---

Ellen Sawilowsky has a Ph. D. in physical chemistry from Case Western Reserve University. She has previously published in journals such as the *Journal of Physical Chemistry* and *Abstracts of the Papers of the American Chemical Society*. Her email address is ell@chemist.com.

$$V = -\frac{q^2\alpha}{2r^4} + \frac{Q}{2} \frac{[(3\cos^2\theta) - 1]}{r^3} \quad (1)$$

where  $r$  is the distance between the ion and the carbon in the  $\text{CO}_2$ ,  $\theta$  is the angle formed by the  $\text{CO}_2$  axis and the line connecting the ion and the carbon atom in the  $\text{CO}_2$ , and  $Q$  is the quadrupole moment.

### Methodology

#### Quantum Mechanical Calculations

The short-range portion of the potential is calculated with the Gaussian 86 suite of programs (Frisch, et al., 1984). MP2 calculations are done using a 6-311++G\*\* basis set. The  $r$  and  $\theta$  values shown in Figure 1 below are varied appropriately. At separation distances ( $r$ 's in Fig. 1) of 1.9 to 6.9 Å, the short-range portion of the potential is a grid of points with spacings every  $15^\circ$  and 0.4 Å connected by a spline fit. Extra data points were added at  $r = 2.3$  Å and  $2.1$  Å and  $\theta = 90^\circ$ ,  $105^\circ$ ,  $120^\circ$ , and  $135^\circ$  and at  $r = 1.9$  Å and  $\theta = 90^\circ$ . The potential energies between the grid points were obtained by means of a cubic spline interpolation (Press, et. al, 1992). These energies are given in Table 1.

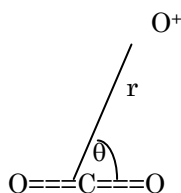


Fig. 1. Parameters used to describe potential surface

#### Long Range Potential

Because quantum chemistry calculations are time-consuming, it is generally more efficient to use classical physics to calculate the potential whenever accuracy allows it. Classical physics gives long-range potential energy terms, which are exact at large separation distances and provide a good analytic form for the long range potential as long as the separation distance is large.

The two potentials which need to be evaluated are the potential which the  $\text{O}^+$  ion induces in the  $\text{CO}_2$  and that which is produced by

the  $\text{CO}_2$ 's charge distribution. The sum of these two potentials provides the analytic form which contains parameters fit to *ab initio* data by minimizing the  $\chi^2$  function:

Table 1. MP2/6-311++G\*\* Energies ( $\text{cm}^{-1}$ )

	$90^\circ$	$105^\circ$	$120^\circ$	$135^\circ$	$150^\circ$	$165^\circ$	$180^\circ$
1.9 Å	3295	-	-	-	-	-	-
2.1 Å	1659	471	492	11363	-	-	-
2.3 Å	1023	-463	-2744	354	-	-	-
2.5 Å	776	-438	-3224	-4121	2833	19351	30093
2.9 Å	627	-89	-2087	-4550	-5457	-3461	-1698
3.3 Å	565	115	-1163	-2981	-4644	-5363	-5391
3.7 Å	496	194	-663	-1899	-3710	-4048	-4332
4.1 Å	421	208	-388	-1251	-2156	-2825	-3067
4.5 Å	349	195	-236	-856	-1511	-2005	-2187
4.9 Å	286	172	-148	-609	-1095	-1465	-1604
5.3 Å	236	150	-94	-447	-818	-1101	-1208
5.7 Å	199	131	-60	-335	-625	-846	-929
6.1 Å	170	115	-38	-258	-487	-662	-727
6.5 Å	147	102	-24	-203	-387	-527	-579
6.9 Å	128	90	-19	-162	-313	-427	-469

$$\chi^2 = \sum_n \left| \frac{V_{\text{fit}} - V_{\text{abinitio}}}{V_{\text{abinitio}}} \right|^2 \quad (2)$$

where  $n$  is the number of points used for the fit,  $V_{\text{fit}}$  is the value of the fitted potential at each point, and the  $V_{\text{ab initio}}$  are the *ab initio* data points used in the fitting process (Bevington & Robinson, 1992). The parameters, which are fit to the *ab initio* points, are the isotropic polarizabilities and the quadrupole moments of  $\text{CO}_2$ . The fit uses the *ab initio* values obtained from Hartree-Fock calculations to begin the parameter search (Levine, 1991). This long-range potential is used to describe the  $\text{CO}_2 + \text{O}^+$  system at separation distances larger than 6.9 Å.

### The Ion-Induced Dipole Term of the Long Range Potential

The ion-induced dipole potential,

$$V(r) = -\frac{q^2 \alpha}{2r^4} \quad (3)$$

where  $q$  is the charge on the ion,  $\alpha$  is the polarizability of the neutral, and  $r$  is the distance between the ion and the center of mass of the neutral, is the potential which the  $O^+$  induces in the  $CO_2$  (Gilbert & Smith, 1990) The polarizability may be expressed as a second order perturbation correction to the dipole moment (Levine, 1991) in a Taylor series expansion of the classical energy of a molecule in the presence of an electric field (Flyglare, 1978).

$$W = W^o + \sum_{\alpha} E_{\alpha} \left( \frac{\partial W}{\partial E_{\alpha}} \right)_{E_{\alpha}=0} + \frac{1}{2} \sum_{\alpha, \beta} E_{\alpha} E_{\beta} \left( \frac{\partial^2 W}{\partial E_{\alpha} \partial E_{\beta}} \right)_{E_{\alpha}=0, E_{\beta}=0} + \dots \quad (4)$$

where  $W$  is the classical potential energy due to the electric field,  $\mathbf{E}$ , and  $\alpha$  and  $\beta$  are the indices for the coordinates.  $\left( \frac{\partial W}{\partial E_{\alpha}} \right)_{E_{\alpha}=0}$  in the first term of

equation 4 is the dipole moment of the molecule and  $\left( \frac{\partial^2 W}{\partial E_{\alpha} \partial E_{\beta}} \right)_{E_{\alpha}=0, E_{\beta}=0}$  in the second term

is the polarizability tensor. In the case of the  $CO_2$  molecule, the dipole moment is zero and the off diagonal elements of the polarizability tensor are zero, reducing equation 4 to the simpler form:

$$W = W^o - \frac{1}{2} \sum_{i=1}^3 \alpha_{ii} E_i^2 \quad (5)$$

The minus sign in Equation 5 is added in order to keep the sign of the polarizability tensor consistent with convention. Because the energy given in Equation 5 is generated from the  $O^+$  point charge, the electric field,  $\vec{E}$ , is given by:

$$\vec{E} = \frac{q}{r^2} \quad (6)$$

where  $q$  is the charge on the ion and  $r$  is the distance between the center of mass of the  $CO_2$  molecule and the  $O^+$  ion. The electric field vector points along the same direction as the vector connecting the center of mass of the  $CO_2$  molecule and the  $O^+$  ion. With  $\theta$  the same angle as shown in the picture in Figure 1, the angle between the line connecting the  $CO_2$  center of mass and the  $O^+$  ion and the line along the body of the  $CO_2$  molecule, the components of the electric field vector areas follows, for a system lying in the  $x$ - $z$  plane:

$$E_x = \frac{q}{r^2} \sin \vartheta \quad (7)$$

$$E_z = \frac{q}{r^2} \cos \vartheta \quad (8)$$

and the second derivative term in (5) becomes:

$$W^{(2)} = -\frac{1}{2} \frac{q^2}{r^4} \left( \alpha_{xx} \sin^2 \vartheta + \alpha_{zz} \cos^2 \vartheta \right) \quad (9)$$

Comparing Equation 9 with Equation 3, it is clear that  $\left( \alpha_{xx} \sin^2 \vartheta + \alpha_{zz} \cos^2 \vartheta \right)$  is the anisotropic form of the polarizability,  $\alpha$  in Equation 3. Equation 9 is the form of the ion-induced dipole potential used in the program that fits the anisotropic polarizabilities to the *ab initio* data. The initial values in the fitting program are the quantum mechanical ones generated from MP2/6-311<sup>++</sup>G\*\* calculations shown in Table 2.

In carrying out the fit, it is important to use the anisotropic form of the polarizability since otherwise all of the angular dependence of the long range potential is in the quadrupole term, giving it a physically unrealistic value, and possibly affecting the accuracy of the potential.

Table 2. Parameters for the Long Range Potential

	<i>Ab initio</i>	Fitted
$\alpha_{xx}$ ( $\text{\AA}^3$ )	1.85	1.68
$\alpha_{zz}$ ( $\text{\AA}^3$ )	3.24	3.68
$\Theta_{xx}$ (Debye- $\text{\AA}$ )	-12.12	-11.89
$\Theta_{zz}$ (Debye- $\text{\AA}$ )	-15.95	-16.53

Note: *ab initio* calculations are done at the MP2/6-311<sup>++</sup>G\*\* level

### Quadrupole Term of the Long Range Potential

The other term of the long range potential is derived from the potential generated by the CO<sub>2</sub> molecule. The potential generated by a collection of charges,  $q_\alpha$ , at a point outside of the body of charges can be expressed as a Taylor series expansion

$$\Phi = \sum_{\alpha} \frac{q_{\alpha}}{r} + \sum_{\alpha,i} q_{\alpha} x_{\alpha,i} \frac{\partial}{\partial x_i} \left( \frac{1}{r} \right) + \frac{1}{2} q_{\alpha} \sum_{i,j} x_{\alpha,i} x_{\alpha,j} \frac{\partial^2}{\partial x_i \partial x_j} \left( \frac{1}{r} \right) + \dots \quad (10)$$

where  $r$  is the distance between the origin and the point and  $x_{\alpha,i}$  is the distance between the origin and the charge  $q_{\alpha}$  (Marion & Heald, 1980). The first term is the monopole term, the second is the dipole term, and the third is the quadrupole term. Although there are several ways to express the quadrupole moment, all of them are based on this third term, which can also be expressed in the form:

$$\Phi^{(3)} = \frac{1}{6} \sum_{i,j} Q_{ij} \frac{(3 x_i x_j - r^2 \delta_{ij})}{r^5} \quad (11)$$

where the  $Q_{ij}$  are components of the quadrupole tensor,  $r$  is the distance from the center of mass of the CO<sub>2</sub> molecule to the ion, and the  $x_i$  are the components of the vector,  $\mathbf{r}$ . This definition of the quadrupole moment is called a traceless quadrupole moment because the trace,  $\sum_i Q_{ii} = 0$ . If the axis along the body of the CO<sub>2</sub> molecule is defined as the  $z$ -axis, and the carbon atom is at the origin, the off-diagonal elements of the quadrupole tensor are zero and  $Q_{xx} = Q_{yy}$ . Because the trace is zero,  $Q_{zz} = -2Q_{xx}$  and there is only one independent element in the quadrupole tensor. Equation 11 becomes:

$$\begin{aligned} \Phi^{(3)} &= \frac{Q_{zz}}{6r^5} \left( -\frac{3}{2}x^2 - \frac{3}{2}y^2 + 3z^2 \right) \\ &= -\frac{Q_{zz}}{4r^5} (r^2 - 3z^2) = \frac{Q_{zz}}{4r^3} (3 \cos^2 \theta - 1) \end{aligned} \quad (12)$$

where  $\theta$  is the angle formed by the line connecting the carbon in CO<sub>2</sub> and the oxygen ion and the  $z$ -axis. The third portion of Equation 12 is the form used for the potential generated by the CO<sub>2</sub> molecule at the location of the oxygen ion.

Quantum mechanical parameters were used instead of experimental ones in the long range potential because (a) a smooth and continuous transition is needed to the short range quantum mechanical potential, and (b) a good comparison between the two is needed in order to decide at what separation distance to change from the long to short range potential. The quantum mechanical quadrupole moments which come out of Gaussian 86 are not the traceless  $Q$ 's in Equation 12, but instead correspond to another definition (Hirschfelder, et al., 1954):

$$\Theta_{ij} = \sum_{\alpha} q_{\alpha} x_{\alpha,i} x_{\alpha,j} \quad (13)$$

where  $q_i$  are the individual charges and  $x_{\alpha,i}$  is the  $i$  component of the vector,  $\mathbf{r}$ , connecting the charge  $\alpha$  to the origin. The analogous traceless definition is (Marion & Heald, 1980):

$$Q_{ij} = \sum_{\alpha} q_{\alpha} (3 x_{\alpha,i} x_{\alpha,j} - r_{\alpha}^2 \delta_{ij}) \quad (14)$$

Substituting equation.13 into 14,

$$Q_{zz} = 3\Theta_{zz} - (\Theta_{xx} + \Theta_{yy} + \Theta_{zz}) \quad (15)$$

and because for the CO<sub>2</sub> molecule,  $\Theta_{xx} = \Theta_{yy}$ ,

$$Q_{zz} = 2(\Theta_{zz} - \Theta_{xx}) \quad (16)$$

hence, Equation 12 becomes:

$$\Phi^{(3)} = \frac{(\Theta_{zz} - \Theta_{xx})}{2r^3} (3 \cos^2 \theta - 1) \quad (17)$$

The potential energy due to the electric field generated by the CO<sub>2</sub> molecule at a point located a distance  $r$  from the carbon is:



$$V = q \frac{(\Theta_{zz} - \Theta_{xx})}{2r^3} (3 \cos^2 \theta - 1) \quad (18)$$

where  $q$  is the charge on the  $O^+$  ion. Equation 18 is the form used in the fitting program and the values for the quadrupole moments,  $\Theta_{zz}$  and  $\Theta_{xx}$ , are generated by Gaussian 86 and given in Table 2.

### Results

#### Combination of Terms to Form the Long-Range Potential

Equations 9 and 18 are added together to give the final form for the long range potential. The two anisotropic polarizability parameters and the two quadrupole moment ones are optimized by doing the  $\chi^2$  fit (Equation 2) to MP2/6-311++G\*\* data points with separation distances of 6.9 Å to 18 Å. Figure 2 shows how the long range potential using the optimized values obtained from the  $\chi^2$  fit compares to the *ab initio* points. The long-range form gives a very accurate representation of the quantum mechanical potential at separation distances larger than 6.9 Å. For this reason, the quantum mechanical grid of points was calculated only for separation distances less than 6.9 Å, and the ion-induced dipole plus quadrupole long range potential was used at larger separation distances. Figure 3 is a contour plot of the entire potential surface.

### Conclusion

It has been demonstrated that a substantial reduction in the amount of time required to produce an accurate potential surface may be obtained by combining the short-range quantum mechanical portion with the less-time intensive long-range one. Starting with an appropriate functional form, the ion-induced dipole and the quadrupole potentials of classical physics, the long-range potential was generated by doing a  $\chi^2$  fit of four parameters to the highly accurate *ab initio* quantum mechanical points. The fitted form of the potential provides the accuracy needed without resorting to difficult quantum mechanical calculations.

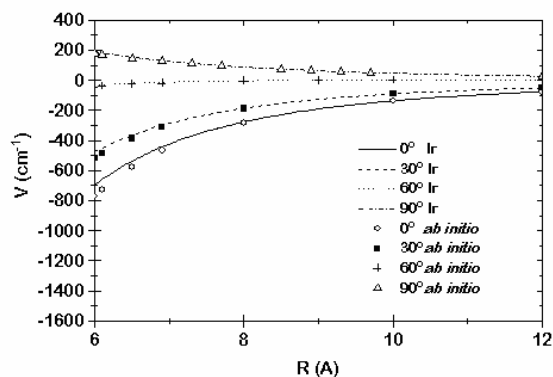


Fig. 2. Comparison of the long range potential with optimized parameters to *ab initio* points.

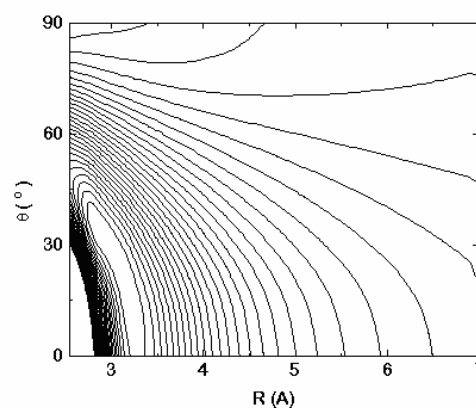


Fig. 3. Contour plot of the complete potential surface for the  $CO_2 + O^+$  system. The contour at the top left corner is  $548 \text{ cm}^{-1}$  and that in the bottom of the well is  $-5328 \text{ cm}^{-1}$ . The contours are spaced  $226 \text{ cm}^{-1}$  apart.

### References

- Bevington, P. R., Robinson, D. K. (1992). *Reduction and error analysis for the physical sciences*. NY: McGraw-Hill, Inc.
- Flyglare, W. H. (1978). *Molecular structure and dynamics*. New Jersey: Prentice-Hall.
- Forst, W. (1973). *Theory of unimolecular reactions*. NY: Wiley-Interscience.

Frisch, M. J., Binkley, J. S., Schlegel, H. B., Raghavachari, K., Melius, C. F., Martin, R. L., Stewart, J. J. P., Bobrowicz, F. W., Rohlfing, C. M., Kahn, L. R., Defrees, D. J., Seeger, R., Whiteside, D. J., Fox, D. J., Fleuder, E. M., & Pople, J. A. (1984). *Gaussian 86*, (Pittsburgh), Carnegie-Mellon Quantum Chemistry.

Gilbert, R. G., Smith, S. C. (1990) *Theory of unimolecular and recombination reactions*. Oxford, England: Blackwell Scientific Publications.

Hirschfelder, J. O., Curtiss, C. F., Bird, R. B. (1954). *Molecular theory of gases and liquids*. NY: John Wiley & Sons, Inc.

Levine, I. N. (1991). *Quantum chemistry*. New Jersey: Englewood Cliffs, Prentice Hall.

Marion, J. B., Heald, M. A. (1980). *Classical electromagnetic radiation*. New York, New York: Academic Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in fortran: the art of scientific programming*. (2<sup>nd</sup> ed.). Cambridge, England: Cambridge University Press.

Viggiano, A. A., Morris, R. A., Van Doren, J. M., & Paulson, J. F. (1992). Temperature, Kinetic Energy, and Internal Energy Dependences of the Rate Constant and Branching Fraction for the Reaction of  $O^+(^4S)$  with  $CO_2$ . *J. Chem. Phys.*, 96, 270.

## A Longitudinal Follow-up Of Discrete Mass At Zero With Gap

Joseph L. Musial  
Department of Internal Medicine  
Henry Ford Health System

Patrick D. Bridge  
Department of Family Medicine  
Wayne State University

Nicol R. Shamey  
Plymouth High School  
Canton, Michigan

---

The first part of this paper discusses a five-year systematic review of the *Journal of Consulting and Clinical Psychology* following the landmark power study conducted by Sawilowsky and Hillman (1992). The second part discusses a five-year longitudinal follow-up of a radically nonnormal population distribution: discrete mass at zero with gap. This distribution was based upon a real dataset.

Key words: Discrete mass at zero with gap, longitudinal data, nonnormality, onset variables, power.

---

### Introduction

There has been a historical concern among researchers and statisticians regarding the prevalence of normally distributed data in real-world populations (Pearson 1895; Geary 1947; Pearson & Please, 1975; Micceri, 1989). For example, Micceri (1989) conducted a study involving population characteristics by examining 440 large-sampled achievement and psychometric data sets in the fields of education and psychology. All of the distributions failed tests of normality, and only 3% remotely resembled a Gaussian distribution (e.g., symmetric with light tails). The concern about nonnormality in real-world data sets has fostered inquiry into the power and robustness of commonly employed parametric statistics under nonnormal conditions (Blair & Higgins, 1980;

Sawilowsky & Hillman, 1992; Sawilowsky & Blair, 1992; Bridge & Sawilowsky, 1999).

An implication of normality is that the probabilities associated with hypothesis tests become inaccurate, and power tables become inexact. Sawilowsky and Hillman (1992) conducted a study that examined the utility of Cohen's (1988) power tables with radically nonnormal distributions. Specifically, the Type I and Type II error properties of the discrete mass at zero distribution were analyzed.

This distribution occurs when portions of the scores fall on zero, and the remaining scores begin to form the shape of the group's distribution. It is common in the fields of public health, as well as education and psychology, and is most prevalent with first use or onset variables, including the age of first cigarette use, age of first alcoholic drink, or the age of first suicide attempt. Sawilowsky and Hillman made two major findings. First, the independent samples  $t$  test was robust as it pertained to Type I error. Second, and thusly, researchers were not discouraged from using Cohen's power tables when analyzing radically nonnormal distributions.

In addition to the findings by Sawilowsky and Hillman (1992), a question was raised regarding the comparative power of radically nonnormal distributions, such as discrete mass at zero with gap. For example, Bridge and Sawilowsky (1999) found the Wilcoxon Rank-Sum test to be more powerful than the independent

---

Joseph L. Musial, Ph.D., is the Education Specialist for the Department of Internal Medicine, Henry Ford Health System, 2799 West Grand Blvd, CFP-1, Detroit, MI 48202-2689. E-mail: [jmusial1@hfhs.org](mailto:jmusial1@hfhs.org). Patrick D. Bridge, Ph.D., is an Assistant Professor of Family Medicine at Wayne State University. Nicol R. Shamey, M.A., is an instructor at Plymouth High School in Canton, MI and practicing psychologist. The authors acknowledge James Hutley and Denise Sigworth of Schoolcraft Community College, Livonia, MI, for their technical assistance.

samples  $t$  test when analyzing distributions with heavy tails or extreme skew, including the discrete mass at zero with gap distribution. Therefore researchers should consider the comparative power of nonparametric statistics when choosing procedures.

An important question stemming from Sawilowsky and Hillman (1992) is what happens to the shape of radically nonnormal distributions over time? Equally important is to assess how researchers approached statistical analysis, as well as the comparative power of nonparametric statistics when faced with extreme nonnormal distributions. For example, were the zero scores re-coded, removed, or treated as outliers? The main point is, however, if the data become normal over time, these issues vanish.

#### Purpose of the Study

The seminal power study conducted by Sawilowsky and Hillman (1992), and Bridge and Sawilowsky (1999) should have raised concerns among researchers and statisticians who encounter radically nonnormal distributions, such as discrete mass at zero with gap. The first purpose of this study was to conduct a five-year systematic review of the *Journal of Consulting and Clinical Psychology*, following Sawilowsky and Hillman (1992), to determine the extent to which researchers who encounter discrete mass at zero with gap address the comparative power issues within their studies. The second purpose is to report on a five-year longitudinal analysis of an academic data set meeting discrete mass at zero with gap. The distributions were assessed in order to determine if there was a shift towards normality or to determine if the distributions remained radically nonnormal overtime.

#### Methodology – Part 1

The *Journal of Consulting and Clinical Psychology* was systematically reviewed over a five-year period following the Sawilowsky and Hillman (1992) publication, involving a power study of the independent samples  $t$  test under a radically nonnormal psychometric distribution. Each article was examined in order to identify any study, which had considered discrete mass at zero with gap or without gap within the context of the population distributions and inclusion variables. Any article that had included onset variables or

distributions that appeared to follow discrete mass at zero with and without gap were flagged.

#### Results

The five-year systematic review identified  $n= 44$  studies that met the criteria for discrete mass at zero with gap (see Appendix). There appeared to be no evidence of the term “discrete mass at zero with gap” used by the authors when either plotting or discussing their distributions. Several studies utilized multiple statistical approaches with scores that fell on zero. For example, Farrell and Danish (1993) re-coded scores with zero, Darkes and Goldman (1993) excluded  $n= 148$  participants due to either non-use (zero) and or extreme scores, and Curran, Stice and Chassin (1997) dropped  $n= 74$  families because a child had reported no (zero) individual and or no (zero) peer alcohol use.

Several studies, however, raised concerns about measurement issues and statistical assumptions. For example, Willett and Singer (1993) introduced discrete-time survival analysis, Loeber and Farrington (1994) discussed violations of population normality, and Gardner, Lidz, Mulvey, and Shaw (1996) noted extreme skew and nonnormality with their discrete mass at zero without gap distribution.

#### Methodology – Part 2

The second phase of this study included identifying a real-live, academic data set which consisted of  $N= 357$  undergraduates who had enrolled in a developmental math course during the Fall of 1995. This cohort was selected because 69 of the students (19%) received a zero in the remedial math course. Each of the students' grade point average (G.P.A.) during the Fall semester was then tracked over a five-year period (1996-2000) in order to describe and analyze the distributions. The academic data were obtained by permission from a mid-western junior college. The appropriate Institutional Review Board approved the study design. All student identifiers were removed from the database and were replaced by a unique identifier.

The cohort was obtained from the colleges' database, with assistance from the school's Institutional Research Office using Microsoft Access 2000 (Microsoft, 2000). The abstracted variables included the developmental math grade for the Fall of 1995, the Fall semester

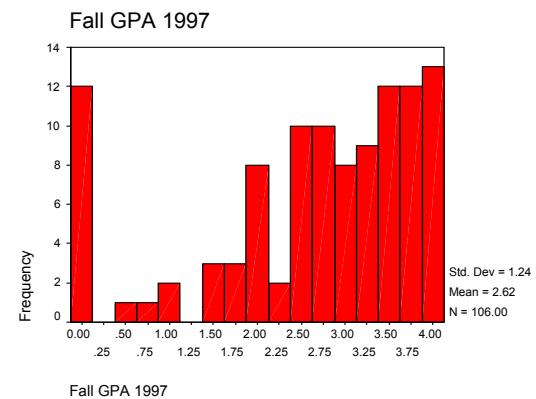
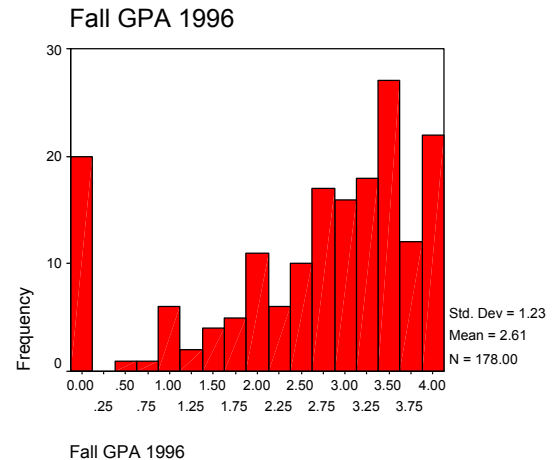
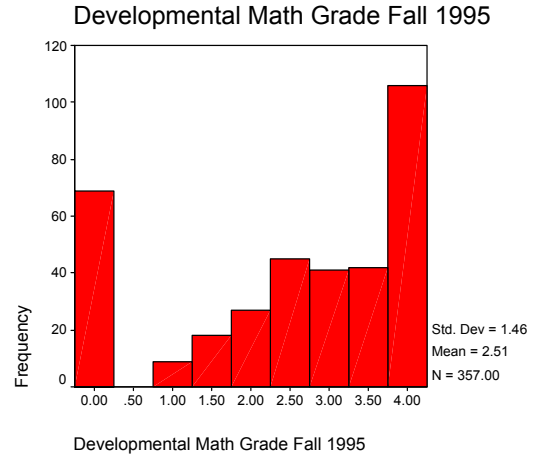
G.P.A. (1996-2000), as well as the unique identifier. The data were then imported into a database using SPSS for Windows, version 11.00 (SPSS Inc, 1999). Descriptive statistics were then generated and included the mean, median, standard deviation, proportions, frequency counts, kurtosis and skew.

Results

Table 1 includes descriptive data derived from the academic distributions. There were a total of n= 69 (19.3%) cases that fell on zero at baseline. This number decreased to n= 4 (1.1%) cases by year 2000. All of the distributions had negative skew and negative kurtosis. Further, all of the distributions remained radically nonnormal over time (see Figure 1 to the right, and continuing on next page). Each distribution could be described as discrete mass at zero with gap except for year 1999, which had no gap. A total of 21 (5.88%) zero scoring performers at baseline had shifted to a positive score at least one time. Additionally, 26 (7.28%) of positive grades at baseline had shifted back to zero at least one time.

Table 1: Descriptive Data

	Base-					
	Line	1996	1997	1998	1999	2000
N	357	178	106	57	44	47
Mean	2.51	2.61	2.62	2.71	2.53	2.76
SD	1.46	1.23	1.24	1.22	1.43	1.20
Skew	-.685	-.991	-.962	-.962	-.650	-1.058
Kurtosis	-.924	-.090	-.080	-.031	-1.001	.277
Scores of Zero						
n	69	20	12	5	6	4
%	19.3	5.6	3.4	1.4	1.7	1.1



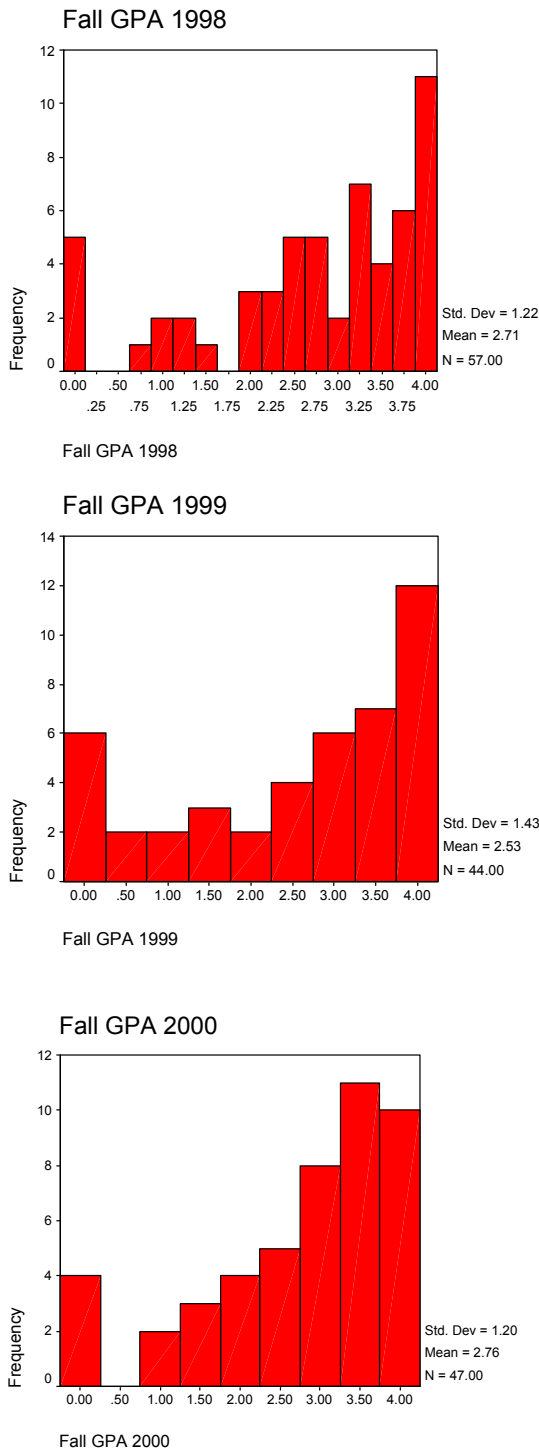


Figure 1 (continued). Distributions.

Conclusion

A systematic five-year review of the *Journal of Consulting and Clinical Psychology* following the

Sawilowsky and Hillman (1992) power publication involving prevalent psychometric distributions with the independent samples *t* test was performed. The results found that none of the authors had considered the outcomes and recommendations reported by Sawilowsky and Hillman despite employing onset variables, which may include radically nonnormal distributions such as discrete mass at zero with gap. This may lead to the inappropriate application of a statistical test, thus, raising concerns about validity.

The compendium clearly diagrams the various approaches that the authors adopted in order to evaluate the variables including recoding zero to a positive number, excluding non-users (those responses who fell on zero), as well as beginning age of onset at age ten. Several authors, however, raised concerns about nonnormality, extreme skew, and the general lack of longitudinal data beyond one year.

The five-year follow-up of discrete mass at zero with gap data set, which was based upon real, radically nonnormal academic data, found that the shape of the distribution remained unchanged over time. Despite a decrease in population size from baseline of N= 357 to N= 47 by year five, the radically nonnormal distribution did not shift towards normality. Four of the five distributions met the criteria for discrete mass at zero with gap, and one distribution, the Fall of 1999, could be described as discrete mass at zero without gap.

An interesting finding among the student G.P.A. scores included the shift from a positive G.P.A. to a zero G.P.A. n= 26 (7.28%) and, vice versa, a shift from a zero G.P.A. to a positive G.P.A. n= 21 (5.88%). This phenomenon may occur with other onset variables, perhaps within a 30-day, 6-month, and 12-month alcoholic relapse log that a family maintains following a loved one's discharge from an inpatient treatment program. However, onset variables such as age at first abortion and or age at first sexual experience do not permit the responder to migrate from a positive value back to a zero response.

Besides understanding onset variables, applied researchers should consider the following three points when analyzing radically nonnormal distributions: 1) Type I error rates are fine and do not make much difference as it relates to power; 2) Researchers are encouraged to use Cohen's (1988)

power tables with no adverse effect; and, 3) A study is likely to have more power if a nonparametric statistic is employed rather than a parametric statistic.

This study represents the first longitudinal report of discrete mass at zero with gap. Future research should investigate other constructs and onset variables in order to determine if the population distributions behave in a similar or dissimilar fashion. It would also be important to gain an understanding of academic data sets in which student scores consistently remain at zero over time as well as to understand the factors associated with migration towards zero.

#### References

- Agras, W. S., Telch, C. F., Arnow, B., Eldredge, K., & Marnell, M. (1997). One-year follow-up of cognitive-behavioral therapy for obese individuals with binge eating disorder. *Journal of Consulting and Clinical Psychology, 65*, 343-347.
- Ball, S., Carroll, K., Babor, T., & Rounsaville, B. (1995). Subtypes of cocaine abusers: Support for a type-A type-B distinction. *Journal of Consulting and Clinical Psychology, 63*, 115-124.
- Barkley, R. A., Guerremont, D. C., Anstopoulos, A. D., & Fletcher, K. E. (1992). A comparison of three family therapy programs for treating family conflicts in adolescents with attention-deficit hyperactivity disorder. *Journal of Consulting and Clinical Psychology, 60*, 450-462.
- Bartlett, S. J., Wadden, T. A., & Vogt, R. A. (1996). Psychosocial consequences of weight cycling. *Journal of Consulting and Clinical Psychology, 64*, 587-592.
- Basen-Engquist, K., & Edmundson, E. W. (1996). Structure of health risk behavior among high school students. *Journal of Consulting and Clinical Psychology, 64*, 764-775.
- Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of the Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. *Journal of Educational Statistics, 5*, 309-335.
- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood-Cliffs, NJ: Prentice Hall.
- Bridge, P. D., & Sawilowsky, S. S., (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and Wilcoxon Rank-Sum Test in small samples applied research. *Journal of Clinical Epidemiology, 52*, 229-235.
- Burman, B., Margolin G., & John, R. S. (1993). America's angriest home videos: Behavioral contingencies observed in home reenactments of marital conflict. *Journal of Consulting and Clinical Psychology, 61*, 28-39.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Hillsdale, N.J.:Erlbaum.
- Curran, P. J., Stice, E., & Chassin, L. (1997). The relation between adolescent alcohol use and peer alcohol use: A longitudinal random coefficients model. *Journal of Consulting and Clinical Psychology, 65*, 130-140.
- Curry, S., McBride, C., Grothaus, L., Loieue, Doug., & Wagner, E. (1995). A randomized trial of self-help materials, personalized feedback and telephone counseling with non-volunteer smokers. *Journal of Consulting and Clinical Psychology, 63*, 1005-1014.
- Darkes, J., & Goldman, M. (1993). Expectancy challenge and drinking reduction: Experimental evidence for a mediational process. *Journal of Consulting and Clinical Psychology, 61*, 344-353.
- Delucchi, K. (1994). Methods for the analysis of binary outcome results in the presence of missing data. *Journal of Consulting and Clinical Psychology, 62*, 569-575.
- Dobkin, P. L., Tremblay, R. E., & Sacchitelle, C. (1997). Predicting boys' early-onset substance abuse from father's alcoholism, son's disruptiveness, and mother's parenting behavior. *Journal of Consulting and Clinical Psychology, 65*, 86-92.
- Domenico, D., & Windle, M. (1993). Intrapersonal and interpersonal functioning among middle-aged female adult children of alcoholics. *Journal of Consulting and Clinical Psychology, 61*, 659-666.
- Drummond, D. C., & Glautier S. (1994). A controlled trial of cue exposure treatment in alcohol dependence. *Journal of Consulting and Clinical Psychology, 62*, 809-817.

- Epstein, E. E., & McCrady, B. S. (1994). Introduction to the special section: Research on the nature and treatment of alcoholism—does one inform the other? *Journal of Consulting and Clinical Psychology, 62*, 1091-1095.
- Fairburn, C. G., Peveler, R. C., Jones, R., Hope, R. A., & Doll, H. (1993). Predictors of 12-month outcome in bulimia nervosa and the influence of attitudes to shape and weight. *Journal of Consulting and Clinical Psychology, 61*, 696-698.
- Farrell, A. D., & Danish, S. J. (1993). Peer drug associations and emotional restraint: Causes or consequences of adolescents' drug use? *Journal of Consulting and Clinical Psychology, 61*, 327-334.
- Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illness. *Journal of Consulting and Clinical Psychology, 64*, 602-609.
- Geary, R. C. (1947). Testing for Normality. *Biometrika, 34*, 209-242.
- Grilo, C. M., Walker, M. L., Becker, D. F., Edell, W. S., & McGlashan, T. H. (1997). Personality disorders in adolescents with major depression, substance use disorders, and coexisting major depression and substance use disorders. *Journal of Consulting and Clinical Psychology, 65*, 328-332.
- Harris, G., Marnie, R., & Quinsey, V. (1994). Psychopathy as a taxon: Evidence that psychopaths are a discrete class. *Journal of Consulting and Clinical Psychology, 62*, 387-397.
- Hiss, H., Foa, E., & Kozak, M. (1994). Relapse prevention for treatment of obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology, 62*, 801-808.
- Hughes, J. (1993). Pharmacotherapy for smoking cessation: Unvalidated assumptions, anomalies, and suggestions for future research. *Journal of Consulting and Clinical Psychology, 61*, 751-760.
- Ichiyama, M. A., & Zucker, R. A. (1996). Articulating subtype differences in self and relational experience among alcoholic men using structural analysis of social behavior. *Journal of Consulting and Clinical Psychology, 64*, 1245-1254.
- Kalichman, S. C., Russell, R. L., Hunter, T. L., & Sarwer, D. B. (1993). Earvin Magic Johnson's HIV serostatus disclosure: Effects on men's perceptions of AIDS. *Journal of Consulting and Clinical Psychology, 61*, 887-891.
- Killen, J. D., Fortman, S. P., Kraemer, H. C., Varady, A., & Newman, B. (1992). Who will relapse? Symptoms of nicotine dependence predict long-term relapse after smoking cessation. *Journal of Consulting and Clinical Psychology, 60*, 797-801.
- Leaf, R. C., DiGiuseppe, R., Mass, R., & Alington, D. E. (1993). Statistical methods for analyses of incomplete clinical service records: Concurrent use of longitudinal and cross-sectional data. *Journal of Consulting and Clinical Psychology, 61*, 495-505.
- Loeber, R., & Farrington, D. (1994). Problems and solutions in longitudinal and experimental treatment studies of child psychopathology and delinquency. *Journal of Consulting and Clinical Psychology, 62*, 887-900.
- Ludwick-Rosenthal, R., & Neufield, R. (1993). Preparation for undergoing an invasive medical procedure: Interacting effects of information and coping style. *Journal of Consulting and Clinical Psychology, 61*, 156-164.
- McMillen, C., Zuravin, S., & Rideout, G. (1995). Perceived benefit from child sexual abuse. *Journal of Consulting and Clinical Psychology, 63*, 1037-1043.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.
- Microsoft Access (2000). Redman, WA: Microsoft Inc.
- Miller-Johnson, S., Emery, R., Marvin, Clarke, W., Lovinger, & Martin, M. (1994). Parent-child relationships and the management of insulin-dependent diabetes mellitus. *Journal of Consulting and Clinical Psychology, 62*, 603-610.
- Mueser, K. T., Bellack, A. S., & Blanchard, J. J. (1992). Comorbidity of schizophrenia and substance abuse: Implications for treatment. *Journal of Consulting and Clinical Psychology, 60*, 845-856.
- Mulhern, R. K., Ochs, J., & Fairclough, D. (1992). Deterioration of intellect among children surviving leukemia: IQ test changes modify estimates of treatment toxicity. *Journal of Consulting and Clinical Psychology, 60*, 477-480.



- Newman, D. L., Moffitt, T. E., Caspi, A., Magdol, L., & Silva, P. A. (1992). Psychiatric disorder in a birth cohort of young adults: Prevalence, comorbidity, clinical significance, and new case incidences from ages 11 to 21. *Journal of Consulting and Clinical Psychology, 64*, 552-562.
- O'Connor, E. A., Carbonari, J. P., & DiClemente, C. C. (1996). Gender and smoking cessation: A factor structure comparison of processes of change. *Journal of Consulting and Clinical Psychology, 64*, 130-138.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society, Ser.A, 186*, 343-414.
- Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223-241.
- Pianta, R. C., Egeland, B., & Adam, E. K. (1996). Adult attachment classification and self-reported psychiatric symptomatology as assessed by the Minnesota Multiphasic Personality Inventory-2. *Journal of Consulting and Clinical Psychology, 64*, 273-281.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*, 352-360.
- Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology, 60*, 240-243.
- Simons, A. D., & Thase, M. E. (1992). Biological markers, treatment outcome, and 1-year follow-up in endogenous depression: Electroencephalographic sleep studies and response to cognitive therapy. *Journal of Consulting and Clinical Psychology, 60*, 392-401.
- Simons, A. D., Gordon, J., Monroe, S., & Thase, M. (1995). Toward an integration of psychologic, social, and biologic factors in depression: Effects on outcome and course of cognitive therapy. *Journal of Consulting and Clinical Psychology, 63*, 369-377.
- SPSS 11.0 for Windows. (1999). Chicago, IL: SPSS Inc.
- St. Lawrence, J. S. (1993). African-American adolescents' knowledge, health-related attitudes, sexual behavior and contraceptive decisions: Implications for the prevention of adolescent HIV infection. *Journal of Consulting and Clinical Psychology, 61*, 104-112.
- St. Lawrence, J. S., Brasfield, T., Jefferson, K., Alleyne, E., O'Bannon, R., & Shirley, A. (1995). Cognitive-behavioral intervention to reduce African American adolescents' risk for HIV infection. *Journal of Consulting and Clinical Psychology, 63*, 221-237.
- Stephens, R., Roffman, R., & Simpson, E. (1994). Treating adult marijuana dependence: A test of the relapse prevention model. *Journal of Consulting and Clinical Psychology, 62*, 92-99.
- Talcott, G., Fiedler, E., Pascale, R., Klesges, R., Peterson, A., & Johnson, R. (1995). Is weight gain after smoking cessation inevitable? *Journal of Consulting and Clinical Psychology, 63*, 313-316.
- Thackwray, D. E., Smith, M. C., Bodfish, J. W., & Meyers, A. W. (1993). A comparison of behavioral and cognitive-behavioral interventions for bulimia nervosa. *Journal of Consulting and Clinical Psychology, 61*, 639-645.
- Webster-Stratton, C., & Hammond, M. (1997). Treating children with early-onset conduct problems: A comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology, 65*, 93-109.
- Wieczorek, W. F., & Miller, B. A. (1992). Preliminary typology designed for treatment matching of driving-while intoxicated offenders. *Journal of Consulting and Clinical Psychology, 60*, 757-765.
- Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology, 61*, 952-965.

## Appendix. Five-year Systematic Review

Information provided from least current to most recent: Author/Year, Population, Inclusion Variable, DMZ (Discrete Mass at Zero) Consideration.

Simons & Thase (1992), 53 patients with major depression, Age of onset of first depression, No

Barkley et al.(1992), 61 adolescents with ADHD, Age of ADHD onset, No

Mulhern et al.(1992), 49 long-term survivors of childhood leukemia, Age at diagnosis, Age at testing, No

Wieczorek & Miller (1992), 156 convicted-while-intoxicated offenders, Age at first drink, No

Killen et al. (1992), 618 smoking cessation participants, Age began smoking, No

Mueser et al. (1992), Review article, Age at first Hospitalization, No

Burman et al. (1993), Married couples: 17 physically aggressive 15 verbally aggressive 18 withdrawing 15 non-distressed, low-conflict Physical aggression scores, No

St. Lawrence (1993), 195 African-American adolescents, Sexual behavior: Number sexual partners & frequency of un-protected sex in past 6 months; Condom use during first intercourse & frequency of protected & unprotected sex in past 6 months, No

Ludwick-Rosenthal & Neufeld (1993), 72 first-time cardiac catheterization patients, Age at first catheterization, No

Farrell & Danish (1993), 1,256 middle school Students, Frequency of drug use past 30 days & frequency of peers offering alcohol & drugs past 30 days, Zero was removed from the scale and replaced with a "1" = never

Darkes & Goldman (1993), 218 male undergraduates screened for a sample of 70 who drank  $\geq 6$  &  $\leq 40$  servings of alcohol/week, 4-week retrospective consumption record, 148 non-users & extreme drinkers were excluded

Leaf et al. (1993), 820 records from 466 female & 361 male, Retrospective analysis included the General Health Questionnaire used to detect acute case onset of distress, Zero treated as the *best possible mental health*. Scattergram provided

Thackwray et al. (1993), 65 bulimic females in different types of treatment for bulimia nervosa, Six-month follow-up of binge eating & purging frequency, 15-69% of participants were abstinent from binge eating & purging

Domencio & Windle (1992), 616 female adult children of alcoholics and non-alcoholics, Number years married Alcohol use past 30 days Cigarette/marijuana use, No

Fairburn et al. (1993), 75 bulimic patients, Degree of attitudinal disturbance: 0-7, 8-10, & 11-12, No

Hughes (1993), Review of pharmacotherapy of smoking cessation, Abstinence rates, DMZ distribution included

Kalichman et al. (1993), 468 males, HIV-related risk factors, Two risk behaviors moved to zero following disclosure at 17 days

Willett & Singer (1993), Review of discrete-time survival analysis as it pertains to event occurrence, Onset of : Suicide ideation Depression Cocaine relapse, Authors introduce discrete-time survival analysis with real clinical data. DMZ distributions included

Stephens et al. (1994), 161 males & 51 females seeking treatment for marijuana use, Age first marijuana use or age first daily use. Alcohol & drug use past 90 days. Marijuana relapse over 12 months., Included DMZ line graph that plots abstinence post-treatment

Harris et al. (1994), 653 serious criminal Offenders, Year of index offense Teen alcohol abuse 0(none) Elementary school maladjustment 0 (never drank), DMZ distributions generated using PCL-R scores

Delucchi (1994), Review of binary outcome results, 2-group  $p$  values, DMZ distributions generated using  $p$  values

Miller-Johnson (1994), 88 children with Type II diabetes, Age at Diagnosis, No

Hiss et al. (1994), 18 participants with obsessive-compulsive disorder, Mean age of onset of symptoms, No

Drummond & Glautier (1994), 35 alcoholic men, Age of first drink. Age first problem drinking. Age first morning drinking. Age first morning withdrawal. Alcohol consumption post follow-up period., No

Loeber & Farrington (1994), Review, Age of onset. Age at termination. Age at committing behavior for the last time., Discussed violations of normality. Notes that it is rare to follow subjects > 1 year.

Epstein & McCrady (1994), Review & Commentary, Age of onset. Degree of sociopathy., Authors suggest comparing subjects along a continua such as age of onset.

Ball et al. (1995), 399 cocaine abusers, Age at onset of drug abuse. Frequency cocaine use past 30 days., No

St. Lawrence et al. (1995), 246 African American adolescents, Age at first intercourse. Number of sex partners past 12 months. Alcohol & marijuana use past 2 months. Perception of personal HIV risk: 0 (no) to 10 (high-risk) scale., No

Talcot et al. (1995), 332 military recruits, Number months smoking. Percent smoking per day: 0-10, 11-20 & 21+., No

Simons et al. (1995), 53 outpatients prior to cognitive therapy treatment, Age at onset of first depression, No

Curry et al. (1995), 1,137 smokers, Age at smoking onset. Longest previous period of abstinence., No

McMillen et al. (1995), 154 low-income women who were sexually abused as children, Age at first abuse, No

O'Connor et al. (1996), 516 smoking cessation participants, Age of onset of smoking. Number of lifetime quit attempts., No

Pianta et al. (1996), 110 women in second trimester of pregnancy, Number of T  $\geq$  65 elevations range: 0 (44%) to 7 (5%), No

Newman et al. (1996), 961, 21-year-olds from New Zealand's Health & Development Study(DMHDS), Age of onset of mental disorders, Authors did not assess disorders before age 10

Bartlett et al. (1996), 130 obese women, Age of onset of obesity. Age first overweight by 6.8 kg. Number diets lasting < 3 days past year., No

Gardner et al. (1996), 357 pairs of psychiatric Emergency Room Patients, Level of seriousness of violence, DMZ distribution included. Authors note extreme skew & non-normality.

Basen-Engquist et al. (1996), 5,537 high school students, 25 health risk behaviors beginning with zero, No

Ichiyama et al. (1996), 274 men in MSU-UM Longitudinal Study, Onset of alcohol-related difficulties over the life-span, No

Dobkin et al. (1997), 82 mother-son dyads subsampled from 1,037 French-speaking Canadian boys. All Fathers were alcoholic, Early-onset of substance abuse, No

Webster-Stratton & Hammond (1997), 97 children with early-onset conduct problems. Parents: 95 mothers & 71 fathers., Age of onset of conduct problems., No

Curran et al. (1997), 363 Hispanic & Caucasian adolescents, Individual & peer alcohol use, 74 families dropped from study because child reported no individual or peer alcohol use

Grilo et al. (1997), 114 adolescent Psychiatric inpatients, Age at first psychiatric contact & psychiatric hospitalization; number of prior psychiatric hospitalizations, No

Agras et al. (1997), 93 obese women, Age of onset of being overweight and age of onset of binge eating, No

## Exploration Of Distributions Of Ratio Of Partial Sum Of Sample Eigenvalues When All Population Eigenvalues Are The Same

Moonseong Heo

Department of Psychiatry  
Weill Medical College  
Cornell University

---

This paper explores empirically the first two moments of ratio of the partial sum of the first two sample eigenvalues to the sum of all eigenvalues when the population eigenvalues of a covariance matrix are all the same. Estimation of the first two moments can be practically crucial in assessing non-randomness of observed patterns on planar graphical displays based on lower rank approximations of data matrices. For derivation of the moments, exact and large sample asymptotic distributions of the sample ratios are reviewed but neither can be applicable to derivation of the moments. Therefore, I rely on simulations, where data matrices  $\mathbf{X}$  with order  $n \times m$  element-wise independent normal distribution with mean 0 and variance  $\sigma^2$  are assumed, that is,  $\mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{nm})$ , and then derive formulas for estimates of means and standard deviations of the sample ratios within a range of order of the data matrix. The derivations are based on the biplot graphical diagnostic methods proposed by Bradu and Gabriel (1976).

Keywords: Bias, biplot, eigenvalues, multivariate Gaussian; Schönemann-Lingoes-Gower coefficient.

---

### Introduction

Lower rank approximations of data matrices  $\mathbf{X}$  ( $n$  rows for individuals,  $m$  columns for variables) are much used in data analysis. The closeness of their fit to  $\mathbf{X}$  is frequently measured by the ratio of the sum of the first  $s$  ( $< m$ ) eigenvalues of  $l_1^2, l_2^2, \dots, l_s^2$  of  $\mathbf{X}^T \mathbf{X}$  to the total of all the eigenvalues  $l_1^2, l_2^2, \dots, l_m^2$  of  $\mathbf{X}^T \mathbf{X}$ , where  $s$  is the rank of the approximation. In particular, the rank  $s$  is usually chosen to be 2 for planar graphical displays, by which data analysts often want to see if they reveal any patterns in population expectations  $E(\mathbf{X}) = \Xi$  and/or covariance structure.

---

The author is very grateful to Dr. K. Ruben Gabriel for his valuable insights and comments, and to John T. Hutchens for the manuscript preparation. This study was supported in part by NIH grants P30DK26687 and P30MH49762. E-mail: moh2002@med.cornell.edu.

Accordingly, confirmation of such visual assessments is usually based on the quantities of the closeness of the planar displays to the data matrix measured by  $r_{(2)}^2 = (l_1^2 + l_2^2) / \sum_{i=1}^m l_i^2$ . This closeness coefficient is equal to the Schönemann - Lingoes - Gower coefficient  $r_{(2)}^2 = \text{trace} \left\{ (\tilde{\mathbf{X}}^T \mathbf{X} \tilde{\mathbf{X}})^{1/2} \right\} / \|\tilde{\mathbf{X}}\| \|\mathbf{X}\|$  (Gower 1971; Lingoes & Schönemann, 1974) as noted by Heo (1996), where  $\tilde{\mathbf{X}}$  is the Euclidian minimum distance rank 2 approximation of  $\mathbf{X}$ .

It has not been clear, however, how large value of  $r_{(2)}^2$  can play the role of a threshold for signaling non-random patterns on the planar displays, which are not overwhelmed by sampling variations. Furthermore, the threshold will depend on the order of data matrices,  $m$  and  $n$ . First, with respect to dependence on  $m$ ,  $r_{(2)}^2$  has its algebraic minimum  $2/m$  because the sample eigenvalues  $l_1^2, l_2^2, \dots, l_m^2$  are ordered in a

descending manner. Secondly, the larger  $n$ , the less will be sampling variations of the patterns of graphical displays. Therefore, observed patterns on graphical displays with  $r_{(2)}^2 = 0.45$  when  $m = 5$  may be less meaningful than those with  $r_{(2)}^2 = 0.45$  when  $m = 30$  for the same  $n$  — the former is relatively much closer to its minimum. One example of the latter case can be found in the biplot of  $n = 100$  archetypal patients with  $m = 30$  psychiatric variables (Strauss et al., 1979; Heo & Gabriel, 2001), where five distinctive clusters of patients of the same diagnosis within each cluster are displayed well enough to convince a data analyst that the patterns on the biplot may indeed represent patterns of population expectation, despite of the moderate  $r_{(2)}^2 = 0.45$ .

The significance of non-random pattern, however, must be inferred based on a sampling distribution of  $r_{(2)}^2$ . Specifically, if an observed  $r_{(2)}^2$  is above the 95 or 97.5 percentile of the sampling distribution, it may indicate that the pattern on planar displays may not be random and may be revealing patterns of population characteristics. Therefore, to provide such thresholds or critical values, I attempt to draw the sampling distribution of  $r_{(2)}^2$  under an  $m$ -variate null Gaussian model:

$$\mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{nm}). \tag{1}$$

In this situation, planar displays of  $\tilde{\mathbf{X}}$  show patterns solely due to random noise  $\sigma^2$ , not due to  $E(\mathbf{X}) = \mathbf{\Xi}$ , and all the eigenvalues of  $E(\mathbf{X}^T \mathbf{X})$ ,  $\lambda_1^2, \lambda_2^2, \dots, \lambda_m^2$ , are the same as  $\sigma^2$ .

I review what is known about the exact and asymptotic distribution of the sample eigenvalues  $l_1^2, l_2^2, \dots, l_s^2$  of  $\mathbf{X}^T \mathbf{X}$  under the null Gaussian model (1) and try to derive sampling distributions of  $r_{(2)}^2$  thereof. However, based on this review and to my knowledge, currently existing normal theories do not seem to be either practical or applicable for derivations of the sampling distribution. Therefore, relying on computer simulations under the null model (1), I attempt to derive empirical models for estimates of  $E(r_{(2)}^2)$

and  $SD(r_{(2)}^2)$ , the first two moments, through assessments of a relative bias  $\beta^2 = E(r_{(2)}^2)/(2/m)$  in comparison to the algebraic minimum and its SD. These two moments can provide basis for normal approximations to the sampling distributions and eventually for the thresholds, or the critical values. I use biplot for a model diagnostic tool as demonstrated in Gabriel and Braud (1971). Issues concerning normal approximation, practical meaning of non-random patterns displayed on the planar spaces and a justification of the null model (1) are discussed.

### Methods

#### Exact distribution

When all the population eigenvalues  $\lambda_i^2$  are equal, i.e.  $\lambda_i^2 = \lambda^2$  for all  $i$ , the exact joint distribution of the sample eigenvalues  $l_i^2$  can be expressed as (e.g., James, 1964):

$$f(\underline{l}^2) = \left(\frac{n}{2\lambda^2}\right)^{nm/2} \frac{\pi^{m^2/2}}{\Gamma_m(n/2)\Gamma_m(m/2)} \exp\left(-\frac{n}{2\lambda^2} \sum_i l_i^2\right) \prod_{i=1}^m l_i^{(n-m-1)} \prod_{i<j} (l_i^2 - l_j^2)$$

where  $\underline{l}^2 = (l_1^2, \dots, l_m^2)^T$  and

$$\Gamma_m(\cdot) = \pi^{m(m-1)/4} \prod_{i=1}^m \Gamma(\cdot - (i-1)/2).$$

Based on this, the exact density of  $r_{(2)}^2 = (l_1^2 + l_2^2) / \sum_{i=1}^m l_i^2$  under the null model, can be obtained by using the change of variable technique. Also, Krishnaiah and Waikar (1971) studied the exact marginal distribution of each individual sample eigenvalue, when all the population eigenvalues are equal, by applying Laplace's expansion to the Vandermonde determinant  $\prod_{i<j} (l_i^2 - l_j^2)$ .

Nevertheless, whichever way is used for calculation of the moments of  $r_{(2)}^2$  under the Null Gaussian model (1), the calculation will be very complicated and tedious, even by numerical computations. Therefore, application of asymptotic or approximation theories might be

preferred for a derivation of the sample moments of  $r_{(2)}^2$  as follows.

Asymptotic distributions

Under the assumption of simplicity (or at least two different multiplicities) of the population eigenvalues, asymptotic (representations for) distributions of the sample eigenvalues were extensively discussed in the 1960s and 70s (e.g., Muirhead, 1978). The joint distributions of sample eigenvalues, under that assumption, involve hypergeometric functions expressed in integral representations. On these integrals are focused the approximations, which are basically determined by the maximum values of the integrands involving ‘linkage factors’ of  $(l_i^2 - l_j^2)^{-1}$ . Such approximations are, therefore, inapplicable to the joint (or marginal) asymptotic behaviors of sample eigenvalues when all the population eigenvalues are equal. Hence, the derivation of an asymptotic distribution of  $r_{(2)}^2$  under multiplicity from the asymptotic joint distribution of sample eigenvalues under the simplicity would be misleading. The following are such examples.

An asymptotic distribution of  $l_i^2$  is  $l_i^2 \sim N(\lambda_i^2, 2\lambda_i^4/(n-1))$  (Anderson, 1963) and  $Cov(l_i^2, l_j^2) \approx 0$  for  $i \neq j$ , provided all eigenvalues are distinct. Under the Null Gaussian model (1), it might become  $l_i^2 \sim N(\sigma^2, 2\sigma^2/(n-1))$  for all  $i$ , if the multiplicity of  $\lambda_i^2$  is ignored, i.e., the fact that  $\lambda_i^2 = \lambda^2$  for all  $i$  is ignored. Applying Taylor approximation to each  $l_i^2$  about each corresponding  $\lambda_i^2$ :

$$\frac{l_1^2 + l_2^2}{\sum l_i^2} = \frac{\lambda_1^2 + \lambda_2^2}{\sum \lambda_i^2} + \frac{1}{\sum \lambda_i^2} \sum (l_i^2 - \lambda_i^2) \left[ I\{i \leq 2\} - \frac{\lambda_1^2 + \lambda_2^2}{\sum \lambda_i^2} \right]$$

where  $I\{\cdot\}$  is an indicator function. Under the Null Gaussian model (1), the right hand side can be reduced to  $2/m + (l_1^2 + l_2^2)/m\sigma^2 - 2\sum l_i^2/m^2\sigma^2$ , which is asymptotically Gaussian with mean  $2/m$  and variance  $4(m-2)/(n-1)m^3$ . This shows very

roughly that the distribution of  $r_{(2)}^2$  does not depend asymptotically on  $\sigma^2$ , as it should not, because  $r_{(2)}^2$  is a studentized ratio. However, the asymptotic expectation  $2/m$  is wrong, since  $(l_1^2 + l_2^2)/\sum l_i^2$  is greater than  $2/m$  with probability one because the sample eigenvalues  $l_i^2$  are ordered in a descending manner.

Asymptotic distributions of functions of sample eigenvalues were investigated by several authors (e.g., Fang & Krishiniah, 1982). Fujikoshi (1980), for example, showed that the distribution functions of functions of sample eigenvalues can be expanded up to the order of  $n^{-1/2}$ , when certain assumptions (including the simplicity of the population eigenvalues) are met. Based on his approximation for the multivariate Gaussian  $\mathbf{X}$ ,  $E(r_{(2)}^2) \approx R_2^2 + a/n$  and  $Var(r_{(2)}^2) = \zeta^2/n$ , where  $R_2^2 = (\lambda_1^2 + \lambda_2^2)/\sum \lambda_i^2$ ,  $a = \sum_{i \neq j} T_i (\lambda_i^2 - \lambda_j^2)^{-1} \lambda_i^2 \lambda_j^2 + \sum T_{ii} \lambda_i^4$ ,  $T_i = I\{i \leq 2 - R_2^2\} / \sum \lambda_i^2$ ,  $T_{ij} = -(T_i + T_j) / \sum \lambda_i^2$ , and  $\zeta^2 = 2 \sum T_i^2 \lambda_i^4$ . Then, apply Fujikoshi's approximations to the set of population eigenvalues such that  $\lambda_i^2 = \lambda_{i+1}^2 + \varepsilon$ , for  $i = 1, \dots, m-1$ , and  $\lambda_m^2 = 1$ , where the difference  $\varepsilon$  of the consecutive population eigenvalues is very small. Numerical evaluations of the expectation of  $r_{(2)}^2 / R_2^2$  and its standard deviation are tabulated in Table 1 for  $\varepsilon = 0.001$ . It is clear from this table that the approximation formulae do not work for these settings of population eigenvalues.

It follows that either exact or asymptotic normal theory does not seem to be applicable to the case of equal eigenvalues. This inapplicability leads us to simulation-based studies, which are described in the following, for empirical exploration of the behavior of the expectation and SD of  $r_{(2)}^2$  under the null Gaussian model (1).

Results

Bias, standard deviation, and simulation fit

The  $n$ -by- $m$  data matrices  $\mathbf{X}$  with  $3 \leq m \leq 30$  and  $30 \leq n \leq 1000$  ( $m \leq n$ ) under the null Gaussian model (1) are randomly generated for 1000 times for each combination of  $n$  and  $m$ , and then  $r_{(2)}^2$  is computed for each data matrix  $\mathbf{X}$ .

Table 1: Asymptotic expectation and (SD) of  $r_{(2)}^2/R_2^2$ :  $\varepsilon=0.001$ .

$n$	$M$					
	3	5	10	15	20	30
30	26.0 (0.1 1)	49.7 (0.1 4)	77.1 (0.1 7)	92.3 (0.1 7)	102. 9 (0.1 7)	118. 1 (0.1 8)
60	13.5 (0.0 7)	25.3 (0.1 0)	39.1 (0.1 2)	46.7 (0.1 2)	52.0 (0.1 2)	59.6 (0.1 3)
90	9.3 (0.0 6)	17.2 (0.0 8)	26.4 (0.0 9)	31.4 (0.1 0)	35.0 (0.1 0)	40.1 (0.1 0)
120	7.3 (0.0 5)	13.2 (0.0 7)	20.0 (0.0 8)	23.8 (0.0 9)	26.5 (0.0 9)	30.3 (0.0 9)
150	6.0 (0.0 5)	10.7 (0.0 6)	16.3 (0.0 7)	19.3 (0.0 8)	21.4 (0.0 8)	24.4 (0.0 8)
500	2.5 (0.0 3)	3.9 (0.0 4)	5.6 (0.0 4)	6.5 (0.0 4)	7.1 (0.0 4)	8.0 (0.0 4)
1000	1.8 (0.0 2)	2.5 (0.0 2)	3.3 (0.0 3)	3.7 (0.0 3)	4.1 (0.0 3)	4.5 (0.0 3)

The sample bias  $B^2$  of  $r_{(2)}^2$  is calculated for each data matrix  $\mathbf{X}$  of the same order by the ratio to its absolute lower bound  $2/m$ , that is,  $B^2 = mr_{(2)}^2/2$ . Table 2 contains averages of  $B^2$  and standard deviations  $SD(B^2)$  from 1,000 simulations for each combination of  $m$  and  $n$ .

Table 2: Averages and (SD) of  $B^2$  from 1000 simulations for each combination of  $m$  and  $n$ .

$n$	$M$					
	3	5	10	15	20	30
30	1.19 (0.0 6)	1.46 (0.0 9)	1.96 (0.1 3)	2.38 (0.1 5)	2.75 (0.1 6)	3.43 (0.1 8)
60	1.13 (0.0 4)	1.32 (0.0 7)	1.65 (0.0 9)	1.92 (0.1 0)	2.16 (0.1 0)	2.58 (0.1 2)
90	1.11 (0.0 4)	1.26 (0.0 5)	1.52 (0.0 7)	1.73 (0.0 8)	1.92 (0.0 8)	2.23 (0.0 8)
120	1.09 (0.0 3)	1.22 (0.0 5)	1.45 (0.0 6)	1.62 (0.0 7)	1.78 (0.0 7)	2.04 (0.0 7)
150	1.08 (0.0 3)	1.20 (0.0 4)	1.40 (0.0 5)	1.55 (0.0 6)	1.68 (0.0 6)	1.92 (0.0 6)
500	1.05 (0.0 2)	1.11 (0.0 2)	1.21 (0.0 3)	1.29 (0.0 3)	1.36 (0.0 3)	1.47 (0.0 3)
1000	1.03 (0.0 1)	1.08 (0.0 2)	1.15 (0.0 2)	1.20 (0.0 2)	1.25 (0.0 2)	1.32 (0.0 2)

It shows that  $B^2$  seems to converge slowly to 1 as  $n$  increases and that the bias depends on the order of  $\mathbf{X}$ ; it goes down with  $n$  but up with  $m$ .

Fit of bias

I first fit averages of  $B^2$ , an estimate of the expected bias  $\beta^2 = E(r_{(2)}^2)/(2/m)$  by taking  $n$  and  $m$  as factor levels. The biplot is used for a model diagnostic tool (Bradu & Gabriel, 1978). The biplot of the data matrix of the averages of  $B^2$  in Table 2 minus the grand mean of the averages is displayed in Figure 1.

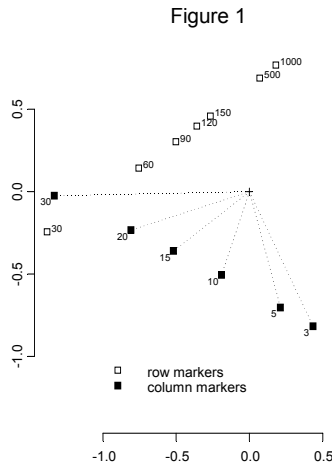


Figure 1: A biplot of  $\beta^2$  with rank 2 goodness of fit greater than 0.99.

This figure shows that the data matrix of  $B^2$  in Table 2 is virtually of rank 2 based on the goodness of fit greater than 0.99. Moreover, it is immediately seen that the sets of column and row markers are both collinear. This suggests that the data matrix must be closely fitted by means of Tukey's Degree of Freedom For Non-Additivity model (DOFNA; Tukey, 1949), i.e.,

$$\beta_{ij}^2 = \mu + \alpha a_i + \delta d_j + \tau a_i d_j + e_{ij} \quad (2)$$

subject to  $\sum a_i = \sum d_j = 0$  and  $\sum a_i^2 = \sum d_j^2 = 1$ . The subscripts  $i$  and  $j$  represent the levels of  $n$  and  $m$ , respectively. (Still, a rank 1 multiplicative model may be an alternative choice. However, a biplot of the data matrix without centering on the grand mean, though not presented herein, shows that the multiplicative model does not fit well.)

A summary graphic of the DOFNA model fit is shown in Figure 2. The residual sum of squares is 0.0037 with df 29, which means that the fit is almost perfect. In short, Figure 2 shows that: (a) There exists a clear interaction between row and column effects, which means that the coefficient  $\tau$  is significantly different from 0:  $\hat{\tau} = 1.84$ ,  $p < 0.001$ ; that is, the magnitude of the bias increases as  $m$  for a fixed  $n$  but the rate of increment is not constant over  $n$ ; (b)  $\beta^2$  seems to converge to 1 as  $n$  increases, as can be seen in Table 2; (c) Roughly, the effect of

the number of columns is close to linear but that of the number of rows is not; the intervals between consecutive row effects are not constant when the magnitude of the number of rows is taken into consideration.

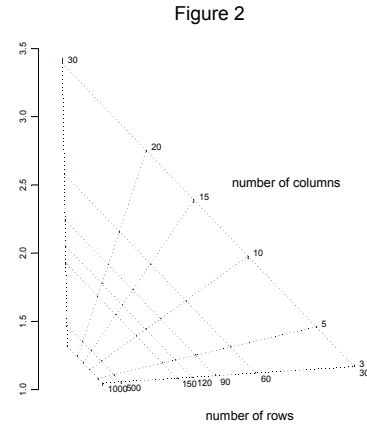


Figure 2: DOFNA fit to  $\beta^2$  with residual sum of squares 0.0037.

It should be recalled, however, that I am trying to formulate a function, which relates this model's parameters to the values (not the factor levels) of  $n$  and  $m$ . For this purpose, on the basis of plots of column effects versus  $m$  and row effects versus  $n$ , we modeled row and column effects as  $\alpha a_i = \gamma_3 / \sqrt{n}$  and  $\delta d_j = \gamma_1 m + \gamma_2 m^2$ , respectively. In light of the DOFNA model (2), this yields the following model:

$$\beta^2 = \eta + \gamma_1 m + \gamma_2 m^2 + (\gamma_3 + \gamma_4 m + \gamma_5 m^2) / \sqrt{n} + e$$

The least-square fit with significant ( $p$ -values  $< 0.001$ ) coefficients results in the following:

$$\hat{\beta}^2 = 1.0301 - 0.0068m + (-0.8319 + 0.6652m - 0.0060m^2) / \sqrt{n} \quad (3)$$

The residual sum of squares of this fit is 0.036 with df 37 and the multiple  $R^2$  is greater than 0.99. All of the fitted values of  $\beta^2$  are greater than



1 over the ranges of  $m$  and  $n$  considered:  $30 \leq n \leq 1000$  and  $3 \leq m \leq 30$ .

Fit of standard deviation

The biplot in Figure 3 with goodness of fit greater than 0.99 shows that the data matrix of  $SD(B^2)$  in Table 2 is also virtually of rank 2 and that the column markers are collinear. On the basis of Bradu and Gabriel (1976), the data matrix of  $SD(B^2)$  must be closely fitted by Mandel's row regression model (Mandel, 1961), that is,

$$SD(B^2) = \mu + \alpha a_i + \delta d_j + \theta c_i d_j + e_{ij}$$

subject to  $\sum a_i = \sum c_i = \sum d_j = 0$  and  $\sum a_i^2 = \sum c_i^2 = \sum d_j^2 = 1$ . The resulting residual sum of squares is  $0.89 \times 10^{-4}$  with df 24, which shows that this is an almost perfect fit.

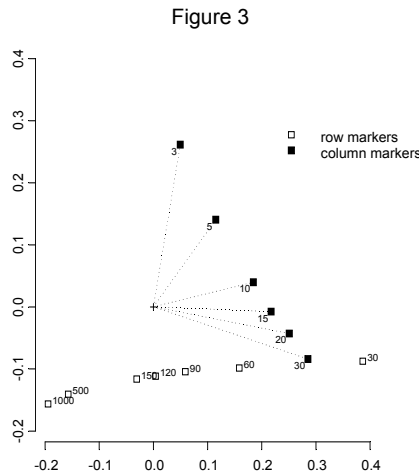


Figure 3: A biplot of  $SD(B^2)$  with rank 2 goodness of fit greater than 0.99.

The biplot in Figure 3, however, shows that the row markers are also virtually collinear. Furthermore, it was observed, though not presented herein, that the  $a_i$ 's and  $c_i$ 's are very similar up to a scale factor. These strongly suggest that Tukey's DOFNA model in a form of (2) can be an alternative fit to the data matrix of  $SD(B^2)$  in Table 2. The DOFNA fit results in a residual sum of squares of  $1.75 \times 10^{-4}$  with df 29. A summary graphic of this DOFNA fit is presented in Figure 4.

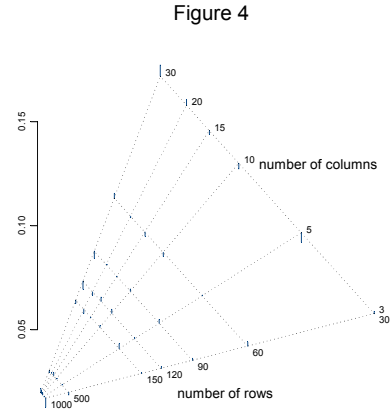


Figure 4: DOFNA fit to  $SD(B^2)$  with residual sum of squares 0.00018.

The structural relationship between  $SD(B^2)$  and the order of  $\mathbf{X}$  is clear;  $SD(B^2)$  seems to vanish slowly as  $n$  increases, which implies  $\beta^2$  converges in probability. Mandel's model is significantly better than the DOFNA model in fitting  $SD(B^2)$  data matrix with an approximated  $F$  ratio 4.65 and  $p$ -value 0.004. This DOFNA model, however, is simpler and easy to see graphically as shown in Figure 4, and its fit is also nearly perfect, which I chose for a functional model construction. Again, based on plots of column effects versus  $m$  and row effects versus  $m$ , I modeled column and row effects as follows:  $\delta d_j = \gamma_1 \log m$  and  $\alpha a_i = \gamma_2 / \sqrt{n}$ , respectively. It follows that

$$SD(B^2) = \eta + \gamma_1 \log m + (\gamma_2 + \gamma_3 \log m) / \sqrt{n} + e$$

(Nevertheless, Mandel's model yields the same form of this model.) The least-square fit with significant ( $p$ -values  $< 0.001$ ) coefficients results in the following:

$$\widehat{SD}(B^2) = 0.0128 - 0.0094 \log m + 0.3123 \log m / \sqrt{n} \tag{4}$$

The residual sum of squares of this fit is  $5.98 \times 10^{-4}$  with df 39 and the multiple  $R^2$  is greater than 0.99. All of the fitted values of

$SD(B^2)$  are positive over the ranges of  $m$  and  $n$  considered:  $30 \leq n \leq 1000$  and  $3 \leq m \leq 30$ .

### Discussion

Regarding features of the distribution of  $r_{(2)}^2$  under the null Gaussian model (1), I observe from the simulation that it is slightly skewed to the right for almost all combinations of  $m$  and  $n$ , but particulars of the asymptotic distributions are unknown. It follows that normal approximation of the distribution of  $r_{(2)}^2$  under the null Gaussian model with the expectation and standard deviation obtained from the formulae (3) and (4) is rather crude. Hypothesis testing based on this normal approximation would, therefore, be conservative. One might consider power transformations of  $r_{(2)}^2$  to have better approximations to normal distributions, or application of “delta” method to the first two moments.

Nevertheless, the crude normal approximation provides an idea of what the distribution of  $r_{(2)}^2$  might be under the null model. For example, to see how many multiples of  $SD(B^2)$  below the mean ensures  $B^2$  to be greater than 1, I calculate a multiple  $c$  from the fitted  $\beta^2$  and  $SD(B^2)$  in the following way:  $c = (\beta^2 - 1)/SD(B^2)$ . From formulae (3) and (4), the estimated minimum  $c$  over the considered ranges is 3.26 when  $m = 3$  and  $n = 30$ . This confirms that  $r_{(2)}^2$  is distributed well above the algebraic minimum of  $2/m$ . Moreover, the multiple  $c$  increases with  $m$ , implying that farther above  $2/m$   $r_{(2)}^2$  is distributed for bigger  $m$ . Indeed, as calculated based on the formulae (3) and (4), the percentiles of  $r_{(2)}^2 = 0.45$  are >99% and 1.4% when  $m$  are 30 and 5, respectively, for the same  $n=100$ . This confirms that observed patterns on graphical displays with  $r_{(2)}^2 = 0.45$  when  $m = 5$  may be less meaningful than those with  $r_{(2)}^2 = 0.45$  when  $m = 30$  for the same  $n$ , as stated in the introduction section.

It has been, however, suspected that  $r_{(2)}^2$  tends to locate between  $\rho_{(2)}^2$  and the absolute

minimum  $2/m$ , where  $\rho_{(2)}^2$  is the “actual” goodness of fit of  $\tilde{\mathbf{X}}$  to the expectation  $\mathbf{\Xi}$ , that is  $\rho_{(2)}^2 = \text{trace}\left\{\left(\tilde{\mathbf{X}}^T \mathbf{\Xi} \mathbf{\Xi}^T \tilde{\mathbf{X}}\right)^{1/2}\right\} / \left\|\tilde{\mathbf{X}}\right\| \left\|\mathbf{\Xi}\right\|$ , which should be a more appropriate measure for the “usefulness” of the lower rank approximation than the measure  $r_{(2)}^2$  of the closeness of  $\tilde{\mathbf{X}}$  to the data  $\mathbf{X}$  themselves, because patterns of the population expectations are to be inferred rather than patterns of data matrix. A simulation study of approximations using data generated under the  $m$ -variate Gaussian model  $\mathbf{X} \sim N(\mathbf{\Xi}, \sigma^2 \mathbf{I}_{nm})$  with affine rank 2 expectation matrix  $\mathbf{\Xi}$  has shown that  $r_{(2)}^2$  indeed underestimates  $\rho_{(2)}^2$  for many situations (Heo and Gabriel, 2001). Thus, non-significant  $r_{(2)}^2$  (less than 95- or 97.5%-tiles of the “null” sampling distribution) implies that the noise  $\sigma$  is much larger relative to the magnitude of  $\mathbf{\Xi}$  — large enough so that  $\|\mathbf{\Xi}\|/\sigma$  is approximately 0. This is the situation where the limiting distribution of  $\mathbf{X} \sim N(\mathbf{\Xi}, \sigma^2 \mathbf{I}_{nm})$  can be approximated by  $\mathbf{X} \sim N(0, \sigma^2 \mathbf{I}_{nm})$  because  $\mathbf{\Xi}$  reaches its zero limit relative to  $\sigma$ . That is, although it maintains all the time its rank, the expectation matrix  $\mathbf{\Xi}$  tends to zero as the magnitude of  $\sigma$  increases, and at the limit it would not have any rank. Therefore, the null distribution  $\mathbf{X} \sim N(0, \sigma^2 \mathbf{I}_{nm})$  is valid for inferences of the critical values for significant  $r_{(2)}^2$ , which indicates that a planar display reveals patterns of population expectation of  $\mathbf{\Xi}$  with a higher  $\rho_{(2)}^2$ .

In sum, the present study shows that there are clear structural patterns of expectation and variance of  $r_{(2)}^2$  under the null Gaussian model (1) as the order of data matrix  $\mathbf{X}$  varies. Construction of formulae for the expectations and standard deviations is elaborated through model diagnosis by use of the biplot. Similar application of the biplot diagnostic method can be extended to exploration of distributions of other ratios of partial sums of sample eigenvalues from data

matrices with bigger orders. The simulation-based approach employed in this paper seems appealing, since any large sample asymptotic theory does not seem to be applicable when all the population eigenvalues are the same. Therefore, the estimated first two moments of  $r_{(2)}^2$  may be useful in judging non-randomness of patterns of population expectations of data matrices displayed in a 2-dimensional space.

#### References

- Anderson, T. W. (1964). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34, 122-148.
- Bradu, D., & Gabriel, K. R. (1978). The biplot as a diagnosis tool for models of two-way tables. *Technometrics*, 20, 46-68.
- Fang, C., & Krishniah, P. R. (1982). Asymptotic distributions of functions of the eigenvalues of some matrices for nonnormal populations. *Journal of Multivariate Analysis*, 12, 39-63.
- Fujikoshi, Y. (1980). Asymptotic expansions for the distributions of the sample roots under nonnormality. *Biometrika*, 67, 45-51.
- Gabriel, K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, 58, 453-467.
- Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In F.R. Hodson, D.G. Kendall, and P. Tautu (Eds), *Mathematics in the Archeological and Historical Sciences*, (pp. 138-149). Edinburgh: University Press.
- Heo, M. (1996). On the fit of sample graphical displays to patterns in population. *Unpublished Dissertation*, University of Rochester.
- Heo, M., & Gabriel, K. R. (2001). The fit of graphical displays to population patterns. *Computational Statistics and Data Analysis*, 36, 47-67.
- James, A. T. (1964). Distributions of matrix variates and latent roots derived from normal samples. *Annals of Mathematical Statistics*, 35, 465-501.
- Krishnaiah, P. R. & Waikar, V. B. (1971). Exact joint distributions of any few ordered roots of a class of random matrices. *Journal of Multivariate Analysis*, 1, 308-315.
- Lingoes, J. C., & Schönemann, P. H. (1974). Alternative measures for fit for the Schönemann-Carroll matrix fitting algorithm. *Psychometrika*, 39, 423-427.
- Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, 56, 878-888.
- Muirhead, R. J. (1978). Latent roots and matrix variates: A review of some asymptotic results. *Annals of Statistics*, 6, 5-33.
- Strauss, J. S., Gabriel, K. R., Kokes, R. F., Ritzler, B. A., VanOrd, A., & Tarana, E. (1979). Do psychiatric patients fit their diagnosis? Patterns of symptomatology as described with the biplot. *Journal of Nervous and Mental Disease*, 167, 105-113.
- Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, 5, 232-242.

## Double Median Ranked Set Sample: Comparing To Other Double Ranked Samples For Mean And Ratio Estimators

Hani M. Samawi  
Department of Mathematics & Statistics  
Sultan Qaboos University  
Sultanate of Oman

Eman M. Tawalbeh  
Department of Statistics  
Yarmouk University  
Jordan

---

Double median ranked set sample (DMRSS) and its properties for estimating the population mean, when the underlying distribution is assumed to be symmetric about its mean, are introduced. Also, the performance of DMRSS with respect to other ranked set samples and double ranked set samples, for estimating the population mean and ratio, is considered. Real data that consist of heights and diameters of 399 trees are used to illustrate the procedure. The analysis and simulation indicate that using DMRSS for estimating the population mean is more efficient than using the other ranked samples and double ranked samples schemes except in case of uniform distribution. Also, using double sampling schemes substantially increase the relative efficiency of ratio estimators relative to their counterpart schemes of one stage samples. Moreover, DMRSS is superior to other double sampling schemes for ratio estimation.

Key words: Double extreme ranked set sample; double median ranked set sample, ratio estimation.

---

### Introduction

In many agricultural and environmental studies and recently in human populations, it is common for quantification of a sampling unit to be costly as compared with the physical acquisition of the unit. For example, level of bilirubin in the blood of infants can be ranked visually by observing: a) color of the face, b) color of the chest, c) color of lower part of the body, & d) color of terminal parts of the whole body. Then, as the yellowish goes from i to iv, the level of bilirubin in the blood goes higher (Samawi & Al-Sakeer, 2001). In such circumstances, considerable cost savings can be achieved if the number of quantification is only a small fraction of the number of available units but all units contribute to the information content of the quantification.

Ranked set sampling (RSS) is considered to be a new method of sampling compared with other sampling methods that can achieve this goal. RSS was first introduced by McIntyre (1952). The use of RSS is highly powerful and much superior to the standard simple random sampling (SRS) for estimating some of the population parameters.

As a variation of RSS Samawi et al. (1996) and Muttlak (1997) investigated extreme ranked set sample (ERSS) and median ranked set sample (MRSS) respectively. Samawi and Muttlak (1996 & 2001) used RSS and MRSS to improve the performance of the ratio estimator. Also, Samawi (2001) suggested the double extreme ranked set sampling (DERSS). They showed that ERSS, MRSS and DERSS are more practical than RSS and more efficient at least than SRS for estimating the population mean. Moreover, Al-Saleh and Al-Kadiri (2000) showed that the efficiency of estimating the population mean could be improved even more by double ranked set sampling technique (DRSS). Also, they proved that ranking in the second stage is easier than in the first stage.

In this article, DMRSS is introduced. The properties of DMRSS for estimating the population mean, when the underlying distribution

---

Hani Michel Samawi is an Associate Professor of Biostatistics. His areas of research are in bootstrap and resampling methods, ranked set sampling estimators, sampling method, testing hypothesis, estimation, and analysis of biostatistics data. E-mail: [hsamawi@squ.edu.om](mailto:hsamawi@squ.edu.om).

is assumed to be symmetric about its mean, are discussed. Also, the performance of DMRSS with respect to the other ranked set samples and double ranked set samples, for estimating the population mean and ratio, is considered.

In Section 2 samples notations and definition and some basic results are introduced . DMRSS scheme and properties are introduced in Section 3. Also, its performance with other sampling schemes will be compared for estimating the population mean. In Section 4, the performance of different double ranked samples schemes will be compared with their counterpart one stage ranked samples for ratio estimation based on the relative efficiency. Illustration of the procedure using real data set with final comments and conclusions is discussed in Section 5.

Sample Notations And Definitions With Some Useful Results  
One Stage Sampling

Univariate population

For any of RSS, ERSS and MRSS schemes, the procedure can be described by selecting r random sets each of size r from the target population. In the most practical situations, the size r will be 2, 3 or 4. Rank each set by a suitable method of ranking like prior information, visual inspection or by the experimenter. In sampling notation this implies:

$$\begin{bmatrix} X_{11}, & X_{12}, & \dots, & X_{1r} \\ X_{21}, & X_{22}, & \dots, & X_{2r} \\ \vdots & & & \vdots \\ X_{r1}, & X_{r2}, & \dots, & X_{rr} \end{bmatrix} \xrightarrow{\text{after ranking}} \begin{bmatrix} X_{1(1)}, & X_{1(2)}, & \dots, & X_{1(r)} \\ X_{2(1)}, & X_{2(2)}, & \dots, & X_{2(r)} \\ \vdots & & & \vdots \\ X_{r(1)}, & X_{r(2)}, & \dots, & X_{r(r)} \end{bmatrix} \tag{2.1}$$

where  $X_{ji}$  denotes the i-th observation in the j-th set and  $X_{j(i)}$  the i-th ordered statistic in the j-th set.

1) If only  $X_{1(1)}, X_{2(2)}, \dots, X_{r(r)}$ , quantified by obtaining the element with smallest rank from the

first set, the second smallest from the second set, and so on until the largest unit from the r-th set is measured. Then, this represents one cycle of RSS. We can repeat the whole procedure m times to get a RSS of size  $n = mr$ . (See Takahasi and Wakimoto, 1968.)

2) Similarly, as in Samawi et al. (1996), we have two cases: In case of r is even, and if only RSS,

$$X_{1(1)k}, X_{2(r)k}, \dots, X_{r-1(1)k}, X_{r(r)k},$$

$k=1,2,\dots,m$ , quantified, then this will denote the  $ERSS_E$ . In case of r is odd, and if only

$$X_{1(1)k}, X_{2(r)k}, \dots, X_{r-1(r)k}, X_{r(\frac{r+1}{2})k},$$

$k=1,2,\dots,m$ , quantified, then this will denote the  $ERSS_O$ .

3) Again, similar to Muttalak (1997), we have two cases: In case of r is odd, and if only

$$X_{1(\frac{r+1}{2})k}, \dots, X_{r(\frac{r+1}{2})k}, k = 1, 2, \dots, m,$$

quantified, then this will denote the  $MRSS_O$ . In case of r is even, select for measurement from the first  $\frac{r}{2}$  samples the  $(\frac{r}{2})$ -th smallest unit and from the last  $\frac{r}{2}$  samples select the  $(\frac{r}{2} + 1)$ -th smallest unit. This will be denoted by  $MRSS_E$  (i.e.

$$X_{1(\frac{r}{2})k}, \dots, X_{\frac{r}{2}(\frac{r}{2})k}, X_{\frac{r}{2}+(\frac{r}{2})k}, \dots, X_{r(\frac{r}{2}+1)k}, k = 1, 2, \dots, m$$

For bivariate population

Samawi and Muttalak (1996) modified the above procedure in case of bivariate distributions to estimate the population ratio. The procedure is described as follows:

First choose  $r^2$  independent bivariate elements from a population, with bivariate distribution function  $F(x, y)$ . Rank each set with respect to one of the variables Y or X. Suppose ranking is on variable X. Apply the same procedures as in case of univariate population but for each measured unit from the X's, the associated unit from the Y's is measured too. This may be repeated m times to get a bivariate sample of size  $n = rm$ .

In sample notation:

1) The sample  $\{(X_{i(i)k}, Y_{i[i]k}), i=1,2,\dots,r, k=1,2,\dots,m\}$  will denote the bivariate RSS.

2) The sample,  $\{(X_{1(1)k}, Y_{1[1]k}), (X_{2(r)k}, Y_{2[r]k}), \dots, (X_{r-1(1)k}, Y_{r-1[1]k}), (X_{r(r)k}, Y_{r[r]k})\}'$

$k=1,2,\dots,m$ , will denote the bivariate ERSS<sub>E</sub> and  $\{(X_{1(1)k}, Y_{1[1]k}), (X_{2(r)k}, Y_{2[r]k}), \dots, (X_{r-1(r)k}, Y_{r-1[r]k}), (X_{r(\frac{r+1}{2})k}, Y_{r[\frac{r+1}{2}]k})\}'$

$k=1,2,\dots,m$ , will denote the bivariate ERSS<sub>O</sub>.

1) Similarly,  $\left\{ \left( X_{i(\frac{r+1}{2})k}, Y_{i[\frac{r+1}{2}]k} \right) : i = 1, 2, \dots, r \right\}$  and  $k = 1, 2, \dots, m$

2) will denote the bivariate MRSS<sub>O</sub> and

$\left( X_{i(\frac{r}{2})k}, Y_{i[\frac{r}{2}]k} \right), \dots, \left( X_{\frac{r}{2}(\frac{r}{2})k}, Y_{\frac{r}{2}[\frac{r}{2}]k} \right)$ ,  $\left( X_{\frac{r}{2}+1(\frac{r+1}{2})k}, Y_{\frac{r}{2}+1[\frac{r+1}{2}]k} \right)$ ,  $\dots, \left( X_{r(\frac{r+1}{2})k}, Y_{r[\frac{r+1}{2}]k} \right)$ ,  $k=1,2, \dots, m$  will denote the bivariate MRSS<sub>E</sub>.

Double Ranked Samples (Two Stage Sampling)

1) Al-Saleh and Al-Kadiri (2000) introduced DRSS procedure as follows:

1. Identify  $r^3$  elements from the target population and divide these elements randomly into  $r$  sets each of size  $r^2$  elements.
2. Use the usual RSS procedure on each set to obtain  $r$  RSS each of size  $r$ .
3. Employ again the RSS procedure in Step 2, to obtain the DRSS of size  $r$ .
4. We may repeat steps 1-3  $m$  times to obtain a sample of size  $n = rm$ .

In sampling notations, after ranking each sample separately in each subset, we get:

$$\begin{bmatrix} X_{1(1)k}^{(1)} & X_{1(2)k}^{(1)} & \dots & X_{1(r)k}^{(1)} \\ X_{2(1)k}^{(1)} & X_{2(2)k}^{(1)} & \dots & X_{2(r)k}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{r(1)k}^{(1)} & X_{r(2)k}^{(1)} & \dots & X_{r(r)k}^{(1)} \end{bmatrix}, \dots, \begin{bmatrix} X_{1(1)k}^{(r)} & X_{1(2)k}^{(r)} & \dots & X_{1(r)k}^{(r)} \\ X_{2(1)k}^{(r)} & X_{2(2)k}^{(r)} & \dots & X_{2(r)k}^{(r)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{r(1)k}^{(r)} & X_{r(2)k}^{(r)} & \dots & X_{r(r)k}^{(r)} \end{bmatrix}, \quad (2.2)$$

$k=1,2,\dots,m$ , where  $X_{i(i)k}^{(l)}$  is the  $i$ -th ordered observation in the  $l$ -th set in the  $i$ -th sample in the  $k$ -th cycle. Use RSS scheme on each subset separately, we get

$$A_{1k} = \{X_{1(1)k}^{(1)}, X_{2(2)k}^{(1)}, \dots, X_{r(r)k}^{(1)}\}, \dots,$$

$$A_{rk} = \{X_{1(1)k}^{(r)}, X_{2(2)k}^{(r)}, \dots, X_{r(r)k}^{(r)}\}$$

Then in the second stage, let  $W_{i(i)k}$  =  $i$ -th smallest observation in  $A_{ik}$ , then  $\{W_{i(i)k}, i=1,2,\dots,r, k=1,2,\dots,m\}$  will denote the DRSS. Now let  $W_{(1)k}, \dots, W_{(r)k}, k=1, 2, \dots, m$ , be a DRSS, with mean and variance of  $W_{(i)k}$  are  $\mu_{(i)}^{**}$  and  $\sigma_{(i)}^{**2}$ , respectively. Al-Saleh and Al-Kadiri (2000) also showed that:

$$\mu = \frac{1}{r} \sum_{i=1}^r \mu_{(i)}^{**} \quad \text{and}$$

$$\sigma^2 = \frac{1}{r} \left[ \sum_{i=1}^r \sigma_{(i)}^{**2} + \sum_{i=1}^r (\mu_{(i)}^{**} - \mu)^2 \right]$$

where  $\mu$  and  $\sigma^2$  are the mean and the variance of the population, respectively. Also, it was shown that ranking in the second stage is easier than in the first stage.

2) DERSS is an extension to ERSS procedure by Samawi (2001). The procedure is just similar to that for DRSS, but taking ERSS instead of RSS in the first and in the second stage. Implies that

$\{W_{1(1)k}, W_{2(r)k}, W_{3(1)k}, \dots, W_{r(r)k}, k = 1, 2, \dots, m\}$  denotes DERSS<sub>E</sub>. The case when  $r$  is odd is similar. For more about RSS see for example Kaur *et al.*, (1995) and Patil *et al.* (1999).

Double Median Ranked Set Sample

In this Section a modification to MRSS, namely double median ranked set sample (DMRSS) is introduced. The properties of this scheme for estimating the population mean, which is considered to be finite, is discussed when the underlying distribution function is assumed to be symmetric. Also, some numerical and theoretical comparisons with SRS, RSS, MRSS, ERSS, DERSS and DRSS are included.

Sample Notation and Definitions

For each cycle  $k=1,2,\dots,m$  ( $m$ = number of cycles), assume a simple random sample, of size  $r^3$ , is selected from a target population with c.d.f.  $F(x)$  and p.d.f.  $f(x)$ , where  $F(x)$  is assumed to be symmetric and absolutely continuous, with mean  $\mu$  and variance  $\sigma^2$ . Suppose we divided the sample independently into  $r$  sets of data where each set contains  $r$  samples, each of size  $r$ . Two cases are considered:

Case 1: From (2.2) and when  $r$  is odd, for the  $k$ -th cycle, get  $r^2$  ranked samples as in (2.2):

Take the median  $X_{i(\frac{r+1}{2})k}^{(j)}$  from each sample in each set, then the following sets are resulted:  $A_{1k} = \{X_{1(\frac{r+1}{2})k}^{(1)}, X_{2(\frac{r+1}{2})k}^{(1)}, \dots, X_{r(\frac{r+1}{2})k}^{(1)}\}$ ,  $A_{2k} = \{X_{1(\frac{r+1}{2})k}^{(2)}, X_{2(\frac{r+1}{2})k}^{(2)}, \dots, X_{r(\frac{r+1}{2})k}^{(2)}\}$ , ...,  $A_{rk} = \{X_{1(\frac{r+1}{2})k}^{(r)}, X_{2(\frac{r+1}{2})k}^{(r)}, \dots, X_{r(\frac{r+1}{2})k}^{(r)}\}$ .

These sets are the first stage MRSS samples. The second stage MRSS or double MRSS is the set of medians of  $A_{1k}, A_{2k}, \dots, A_{rk}$ . Define  $W_{1(\frac{r+1}{2})k} = \text{med}(A_{1k})$ ,  $W_{2(\frac{r+1}{2})k} = \text{med}(A_{2k})$ , ...,  $W_{r(\frac{r+1}{2})k} = \text{med}(A_{rk})$ , then the sample  $\{W_{1(\frac{r+1}{2})k}, W_{2(\frac{r+1}{2})k}, \dots,$

$W_{r(\frac{r+1}{2})k}\}$ ,  $k=1,2,\dots,m$  is denoted by DMRSS<sub>O</sub>.

The sample mean using DMRSS<sub>O</sub> is given by

$$\bar{W}_{DMRSSO} = \frac{1}{rm} \sum_{k=1}^m \sum_{i=1}^r W_{i(\frac{r+1}{2})k} \tag{3.1}$$

Case 2: When  $r$  is even, for the  $k$ -th cycle, after ranking each sample in each set, as in Case 1, divide the  $r$  sets in (2.2) in half to two independent sets. From the first  $\frac{r}{2}$  sets take the  $(\frac{r}{2})$ -th smallest unit from each sample and from the last  $\frac{r}{2}$  sets take the  $(\frac{r}{2} + 1)$ -th smallest unit from each sample, that is, we will get the following sets:

$$A_{1k} = \left\{ X_{1(\frac{r}{2})k}^{(1)}, X_{2(\frac{r}{2})k}^{(1)}, \dots, X_{r(\frac{r}{2})k}^{(1)} \right\},$$

$$A_{2k} = \left\{ X_{1(\frac{r}{2})k}^{(2)}, X_{2(\frac{r}{2})k}^{(2)}, \dots, X_{r(\frac{r}{2})k}^{(2)} \right\}, \dots,$$

$$A_{\frac{r}{2}k} = \left\{ X_{1(\frac{r}{2})k}^{(\frac{r}{2})}, X_{2(\frac{r}{2})k}^{(\frac{r}{2})}, \dots, X_{r(\frac{r}{2})k}^{(\frac{r}{2})} \right\},$$

$$B_{1k} = \left\{ X_{1(\frac{r}{2}+1)k}^{(\frac{r}{2}+1)}, X_{2(\frac{r}{2}+1)k}^{(\frac{r}{2}+1)}, \dots, X_{r(\frac{r}{2}+1)k}^{(\frac{r}{2}+1)} \right\},$$

$$B_{2k} = \left\{ X_{1(\frac{r}{2}+1)k}^{(\frac{r}{2}+2)}, X_{2(\frac{r}{2}+1)k}^{(\frac{r}{2}+2)}, \dots, X_{r(\frac{r}{2}+1)k}^{(\frac{r}{2}+2)} \right\}, \dots,$$

$$B_{\frac{r}{2}k} = \left\{ X_{1(\frac{r}{2}+1)k}^{(r)}, X_{2(\frac{r}{2}+1)k}^{(r)}, \dots, X_{r(\frac{r}{2}+1)k}^{(r)} \right\},$$

$k=1,2,\dots, m$ . This is the first stage. Again from each  $A_{ik}$  take the  $(\frac{r}{2})$ -th smallest units, while from each  $B_{ik}$  take the  $(\frac{r}{2} + 1)$ -th smallest unit as follows:  $W_{i(\frac{r}{2})k} = \text{the } \left(\frac{r}{2}\right)\text{-th ordered statistic from } A_{ik}, i=1, 2, \dots, \frac{r}{2}, k= 1, 2, \dots, m$  and  $W_{i(\frac{r}{2}+1)k} = \text{the } \left(\frac{r}{2} + 1\right)\text{-th ordered statistic from } B_{ik}, i=1, 2, \dots, \frac{r}{2}, k= 1,2,\dots,m$ . Then the resulted

sample  $W_{1(\frac{r}{2})k}, W_{2(\frac{r}{2})k}, \dots, W_{\frac{r}{2}(\frac{r}{2})k},$   
 $W_{1(\frac{r+1}{2})k}, W_{2(\frac{r+1}{2})k}, \dots, W_{\frac{r}{2}(\frac{r+1}{2})k}$ ,  
 $k=1,2,\dots,m$  denotes DMRSS<sub>E</sub>. The sample mean using DMRSS<sub>E</sub> is given by  

$$\bar{W}_{DMRSS E} = \frac{1}{rm} \sum_{k=1}^m \left( \sum_{i=1}^{r/2} W_{i(\frac{r}{2})k} + \sum_{i=1}^{r/2} W_{i(\frac{r+1}{2})k} \right). \tag{3.2}$$

To study the properties of DMRSS<sub>O</sub> and DMRSS<sub>E</sub>, next we derive the distribution functions of  $W_{(\frac{r+1}{2})}, W_{(\frac{r}{2})}$  and  $W_{(\frac{r+1}{2})}$  respectively and some of their properties.

Distribution Function and Properties of DMRSS  
 Case 1: When  $r$  is odd. To find the distribution of  $W_{i(\frac{r+1}{2})k}, i=1,2, \dots, r$  say  $G_{(\frac{r+1}{2})}(w)$ , first the distribution of  $X_{i(\frac{r+1}{2})k}^{(j)}$  say  $F_{(\frac{r+1}{2})}(x)$  is given by

$$F_{(\frac{r+1}{2})}(x) = \int_{-\infty}^x \frac{r!}{\left(\frac{r-1}{2}\right)! \left(\frac{r-1}{2}\right)!} f(t) (F(t))^{\left(\frac{r-1}{2}\right)} (1-F(t))^{\left(\frac{r-1}{2}\right)} dt \tag{3.3}$$

(see Arnold, et al. 1992). Let  $u = F(t)$ , then

$$F_{(\frac{r+1}{2})}(x) = \int_0^{F(x)} \frac{r!}{\left(\frac{r-1}{2}\right)! \left(\frac{r-1}{2}\right)!} (u)^{\left(\frac{r-1}{2}\right)} (1-u)^{\left(\frac{r-1}{2}\right)} du = I_{F(x)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right), \tag{3.4}$$

which is the usual incomplete beta function. Hence,  $X_{1(\frac{r+1}{2})k}^{(j)}, X_{2(\frac{r+1}{2})k}^{(j)}, \dots, X_{r(\frac{r+1}{2})k}^{(j)}$ , are independent and identically distributed (i.i.d.) with incomplete beta  $I_{F(x)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right)$  distribution. Now from the definition of DMRSS<sub>O</sub>, the p.d.f. of  $W_{i(\frac{r+1}{2})k}$  will be

$$g_{(\frac{r+1}{2})}(w) = \frac{r!}{\left(\frac{r-1}{2}\right)! \left(\frac{r-1}{2}\right)!} \left(F_{(\frac{r+1}{2})}(w)\right)^{\left(\frac{r-1}{2}\right)} \left(1-F_{(\frac{r+1}{2})}(w)\right)^{\left(\frac{r-1}{2}\right)} f_{(\frac{r+1}{2})}(w) = \left(\frac{r!}{\left(\frac{r-1}{2}\right)! \left(\frac{r-1}{2}\right)!}\right)^2 \left(\left(I_{F(w)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right)\right)\left(1-I_{F(w)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right)\right)\right)^{\left(\frac{r-1}{2}\right)} \times \left(F(w)(1-F(w))\right)^{\left(\frac{r-1}{2}\right)} \times f(w). \tag{3.5}$$

Let  $s = I_{F(t)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right)$ , then the c.d.f. of  $W_{i(\frac{r+1}{2})k}$  is

$$G_{(\frac{r+1}{2})}(w) = \int_0^{I_{F(w)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right)} \frac{r!}{\left(\frac{r-1}{2}\right)! \left(\frac{r-1}{2}\right)!} (s)^{\left(\frac{r-1}{2}\right)} (1-s)^{\left(\frac{r-1}{2}\right)} ds = I_{I_{F(w)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right)}\left(\frac{r+1}{2}, \frac{r+1}{2}\right). \tag{3.6}$$

Note that,  $W_{1(\frac{r+1}{2})k}, W_{2(\frac{r+1}{2})k}, \dots, W_{r(\frac{r+1}{2})k}, k=1, 2, \dots, m$  are i.i.d. with the (3.6) distribution function.

Case 2: Distribution function of  $W_{i(\frac{r}{2})}$  and  $W_{i(\frac{r+1}{2})}$ ,  $i=1, 2, \dots, \frac{r}{2}$  when  $r$  is even.

Recall the assumption that the MRSS is based on a simple random sample of size  $r^3$  with the symmetric and i.i.d. distribution function  $F(x)$ .

Distribution function of  $W_{i(\frac{r}{2})}$

Using the same steps as in case 1, the p.d.f. and c.d.f. of  $W_{i(\frac{r}{2})k}, i = 1, 2, \dots, \frac{r}{2}$ , will be respectively



$$\begin{aligned}
 g_{\left(\frac{r}{2}\right)}(w) &= \frac{r!}{\left(\frac{r}{2}\right)!\left(\frac{r}{2}-1\right)!} \left(F_{\left(\frac{r}{2}\right)}(w)\right)^{\left(\frac{r}{2}-1\right)} \\
 &\quad \left(1-F_{\left(\frac{r}{2}\right)}(w)\right)^{\left(\frac{r}{2}\right)} f_{\left(\frac{r}{2}\right)}(w) \\
 &= \left(\frac{r!}{\left(\frac{r}{2}\right)!\left(\frac{r}{2}-1\right)!}\right)^2 \left(I_{F(w)}\left(\frac{r}{2}, \frac{r}{2}+1\right)\right)^{\left(\frac{r}{2}-1\right)} \\
 &\quad \left(1-I_{F(w)}\left(\frac{r}{2}, \frac{r}{2}+1\right)\right)^{\left(\frac{r}{2}\right)} \\
 &\quad \times \left(F(w)\right)^{\left(\frac{r}{2}-1\right)} \left(1-F(w)\right)^{\left(\frac{r}{2}\right)} \times f(w)
 \end{aligned}
 \tag{3.7}$$

and

$$\begin{aligned}
 G_{\left(\frac{r}{2}\right)}(w) &= \\
 &\int_0^{I_{F(w)}\left(\frac{r}{2}, \frac{r}{2}+1\right)} \frac{r!}{\left(\frac{r}{2}\right)!\left(\frac{r}{2}-1\right)!} (s)^{\left(\frac{r}{2}-1\right)} (1-s)^{\left(\frac{r}{2}\right)} ds \\
 &= I_{I_{F(w)}\left(\frac{r}{2}, \frac{r}{2}+1\right)}\left(\frac{r}{2}, \frac{r}{2}+1\right) .
 \end{aligned}
 \tag{3.8}$$

Note that, the  $W_{i\left(\frac{r}{2}\right)k}$ ,  $i=1, 2, \dots, \frac{r}{2}$ ,  $k=1, 2, \dots, m$  are i.i.d. with (3.7) distribution function. Similarly, the p.d.f. and c.d.f. of  $W_{i\left(\frac{r}{2}\right)k}$ ,  $i=1, 2, \dots, \frac{r}{2}$ ,  $k=1, 2, \dots, m$ , say,  $g_{\left(\frac{r}{2}+1\right)}(w)$  and  $G_{\left(\frac{r}{2}+1\right)}(w)$  respectively are,

$$\begin{aligned}
 g_{\left(\frac{r}{2}+1\right)}(w) &= \frac{r!}{\left(\frac{r}{2}\right)!\left(\frac{r}{2}-1\right)!} \left(F_{\left(\frac{r}{2}+1\right)}(w)\right)^{\left(\frac{r}{2}\right)} \\
 &\quad \left(1-F_{\left(\frac{r}{2}+1\right)}(w)\right)^{\left(\frac{r}{2}-1\right)} f_{\left(\frac{r}{2}+1\right)}(w) \\
 &= \left(\frac{r!}{\left(\frac{r}{2}\right)!\left(\frac{r}{2}-1\right)!}\right)^2 \left(I_{F(w)}\left(\frac{r}{2}+1, \frac{r}{2}\right)\right)^{\left(\frac{r}{2}\right)} \\
 &\quad \times \left(1-I_{F(w)}\left(\frac{r}{2}+1, \frac{r}{2}\right)\right)^{\left(\frac{r}{2}-1\right)} \\
 &\quad \times \left(F(w)\right)^{\left(\frac{r}{2}\right)} \left(1-F(w)\right)^{\left(\frac{r}{2}-1\right)} f(w)
 \end{aligned}
 \tag{3.9}$$

and,

$$\begin{aligned}
 G_{\left(\frac{r}{2}+1\right)}(w) &= \\
 &\int_0^{I_{F(w)}\left(\frac{r}{2}+1, \frac{r}{2}\right)} \frac{r!}{\left(\frac{r}{2}\right)!\left(\frac{r}{2}-1\right)!} (u)^{\left(\frac{r}{2}\right)} (1-u)^{\left(\frac{r}{2}-1\right)} du \\
 &= I_{I_{F(w)}\left(\frac{r}{2}+1, \frac{r}{2}\right)}\left(\frac{r}{2}+1, \frac{r}{2}\right) .
 \end{aligned}
 \tag{3.10}$$

Hence,  $W_{1\left(\frac{r}{2}+1\right)k}$ ,  $W_{2\left(\frac{r}{2}+1\right)k}$ ,  $\dots$ ,  $W_{\frac{r}{2}\left(\frac{r}{2}+1\right)k}$ , are i.i.d. with the distribution function as in (3.10). However,  $W_{1\left(\frac{r}{2}\right)k}$ ,  $W_{2\left(\frac{r}{2}\right)k}$ ,  $\dots$ ,  $W_{\frac{r}{2}\left(\frac{r}{2}\right)k}$ ,  $W_{1\left(\frac{r}{2}+1\right)k}$ ,  $W_{2\left(\frac{r}{2}+1\right)k}$ ,  $\dots$ ,  $W_{\frac{r}{2}\left(\frac{r}{2}+1\right)k}$ , are independent but not identically distributed.

### DMRSS for Estimating the Population Mean

The following results are stated and proved in the Appendix. Using DMRSS when the underlying distribution is assumed to be symmetric.  $\bar{W}_{DMRSS_O}$  and  $\bar{W}_{DMRSS_E}$  are unbiased estimators for  $\mu$ , and  $Var\left(\bar{W}_{DMRSS}\right) \leq Var\left(\bar{X}_{MRSS}\right) \leq \frac{\sigma^2}{n}$ , where

$$\bar{X}_{MRSS} = \begin{cases} \frac{1}{n} \sum_{k=1}^m \sum_{i=1}^{r/2} (X_{i(\frac{r}{2})k} + X_{i(\frac{r+1}{2})k}), & \text{if } r \text{ is even} \\ \frac{1}{n} \sum_{k=1}^m \sum_{i=1}^r X_{i(\frac{r+1}{2})k}, & \text{if } r \text{ is odd} \end{cases}$$

(see the Appendix for Theorem 1, Lemma 1 and Theorem 2.

Simulation Study

Based on 5000 replication, a computer simulation is conducted to study the behavior of the efficiency of the sample mean using SRS, MRSS, RSS, ERSS, DERSS and DRSS with respect to DMRSS. Random observations are generated from (1) standard normal distribution (2) Logistic distribution with  $\alpha =2, \beta=1$  and (3) uniform distribution with  $\theta_1 =0, \theta_2=4$ . The performance of the samples means for  $r=4,5,6$  and  $7$  and  $m=4$  and  $6$  are investigated.

Results of simulation study

The results of these simulations are summarized by the relative efficiency ( the ratio of the variances) of the estimators of the mean. The simulation results are given in Table 3.1.

Table 3.1 shows that estimating the population mean using DMRSS is substantially more efficient than SRS, MRSS, ERSS and RSS. Comparing the sample mean using MRSS with the sample mean using DMRSS, our simulation confirms the results of Theorem 3.2 for the three distributions. Comparing the efficiency for estimating the population mean using DMRSS relative to DERSS, there is a notable difference between them according to the distributions. The best performance was in case of logistic distribution. In normal distribution, the relative efficiency was slightly lower than in logistic distribution.

Clearly in case of uniform distribution, estimating the population mean using DERSS is more efficient than using DMRSS. Also, the population mean estimator using DMRSS is more efficient than the population mean estimator using DRSS, when the underlying distribution is assumed to be symmetric.

Regarding the sample size  $r$ , the relative efficiency of the population mean estimators, using DMRSS with respect to any of the other

Table 3.1: The efficiency of the mean estimators using DMRSS relative to the others

m	r	SRS	MRSS	ERSS	RSS	DERSS	DRSS
Normal(2,1)							
4	4	7.51	2.74	3.56	3.13	2.71	1.99
	5	11.83	3.49	5.20	4.53	3.69	2.81
	6	16.41	4.16	6.73	5.21	4.59	3.01
	7	23.36	4.88	8.32	6.34	5.98	3.59
6	4	7.28	2.69	3.67	3.18	2.71	2.01
	5	12.42	3.66	5.26	4.67	3.83	2.81
	6	15.85	3.82	6.86	4.83	4.45	2.81
7	22.96	4.93	8.54	6.65	6.13	3.80	
Logistic (2,1)							
4	4	8.53	2.57	5.20	3.79	4.54	2.80
	5	14.63	3.61	7.70	6.06	7.00	4.08
	6	19.38	4.11	10.74	6.54	10.25	4.35
	7	29.47	4.81	14.74	8.79	13.95	5.43
6	4	8.50	2.77	5.11	3.89	4.61	2.79
	5	15.03	3.77	7.28	6.13	6.90	4.07
	6	19.85	3.83	11.68	6.29	9.84	4.11
7	29.86	4.95	14.06	8.88	13.26	5.62	
Uniform(0,4)							
4	4	4.35	2.17	1.44	1.83	0.52	1.00
	5	6.82	2.97	1.89	2.29	0.81	1.21
	6	8.95	3.34	1.70	2.58	0.26	1.27
	7	11.75	4.00	2.10	3.10	0.68	1.35
6	4	4.46	2.20	1.45	1.78	0.50	1.04
	5	7.45	3.11	1.94	2.39	0.87	1.25
	6	8.99	3.28	1.60	2.50	0.27	1.27
	7	18.19	4.40	1.75	3.42	0.16	1.46

previous sampling techniques, increases as  $r$  increases. While considering the cycle size  $m$ , the relative efficiency for the sample mean using DMRSS relative to the other sampling schemes is not affected by the value of  $m$ .

Ratio Estimators

Frequently the quantity that is to be estimated from a bivariate random sample is the ratio of two means of two correlated variables, say  $X$  and  $Y$ , which both vary from unit to unit. For example, in a household survey, the average expenditure on cosmetics per adult female, and the average number of hours per week spent watching television for child aged 10 to 15.

Examples of this kind occur frequently when the sampling unit (the household) comprises a group or cluster of elements and our interest is in the population mean per element. Also, the ratio estimation method is used to obtain increased precision of estimating the population mean or total by taking advantage of the correlation between an auxiliary variable  $X$  and the variable of interest  $Y$ . In this paper, we assume that the

bivariate random variable (X,Y) has symmetric marginal distributions.

#### Ratio Estimator Using SRS

Let the bivariate random variable (X,Y) has c.d.f. F(x,y) with means  $\mu_x$  and  $\mu_y$ , variances  $\sigma_x^2$  and  $\sigma_y^2$ , and correlation coefficient  $\rho$ , then

$R = \frac{\mu_y}{\mu_x}$  will denote the population ratio. Using a

simple bivariate random sample from F(x, y), the estimator of R is given by:

$$\hat{R}_{SRS} = \frac{\bar{Y}}{\bar{X}}$$

(4.1)

where  $\bar{X}$  and  $\bar{Y}$  are the means of X and Y respectively.

Hansen et al.(1953) showed that the variance of  $\hat{R}_{SRS}$  can be approximated by

$$Var(R_{SRS}) \cong \frac{R^2}{n} \left( V_x^2 + V_y^2 - 2\rho V_x V_y \right), \quad (4.2)$$

where  $V_x = \frac{\sigma_x}{\mu_x}$ ,  $V_y = \frac{\sigma_y}{\mu_y}$ ,

$$\rho = \frac{E((X - \mu_x)(Y - \mu_y))}{\sigma_x \sigma_y} \text{ and } n = rm.$$

#### Ratio Estimator Using RSS

Samawi and Muttalak (1996) showed that the ratio estimator using RSS when ranking is on

the variable X, is  $\hat{R}_{RSS2} = \frac{\bar{Y}_{[r]}}{\bar{X}_{(r)}}$ , where

$$\bar{Y}_{[r]} = \frac{1}{n} \sum_{k=1}^m \sum_{i=1}^r Y_{i[i]k}, \quad \bar{X}_{(r)} = \frac{1}{n} \sum_{k=1}^m \sum_{i=1}^r X_{i(i)k} \text{ and}$$

the variance is given by

$$Var(\hat{R}_{RSS}) \cong \frac{R^2}{n} \times \left\{ V_x^2 + V_y^2 - 2\rho V_x V_y \right. \\ \left. \times \left[ \frac{\sum_{i=1}^r T_{x(i)}^2}{n \mu_x^2} + \frac{\sum_{i=1}^r T_{y[i]}^2}{n \mu_y^2} - 2 \frac{\sum_{i=1}^r T_{x(i)y[i]}}{n \mu_x \mu_y} \right] \right\}, \quad (4.3)$$

where  $T_{x(i)} = \mu_{(i)} - \mu_x$ ,  $T_{y[i]} = \mu_{y[i]} - \mu_y$  and

$$T_{x(i)y[i]} = (\mu_{x(i)} - \mu_x)(\mu_{y[i]} - \mu_y).$$

As demonstrated by Samawi and Muttalak (1996), that ranking on X is more efficient than ranking on Y in ratio estimation in terms of variance, therefore we introduce only the case where ranking on the variable X is assumed to be without errors. In the next subsections, we will introduced and study the performance of ratio estimators using the double ranked samples discussed in the pervious sections.

#### Ratio Estimation Using DRSS

Using the notation of Section 2.3, the second stage a subsample of size  $n=rm$ ,  $\{W_{i(i)k}, i = 1, 2, \dots, r, k = 1, 2, \dots, m\}$  is selected.

Also, in the second stage, for each  $W_{i(i)k}$  measure (quantify) the associated value of the random variable Y. The bivariate DRSS  $\{(W_{i(i)k}, Y_{i[i]k}) : i = 1, 2, \dots, r, k = 1, 2, \dots, m\}$  is measured, where  $W_{i(i)k}$  as defined above, and

$Y_{i[i]k}$  is the corresponding value of Y obtained from the i-th RSS sample in the k-th cycle.

Now, let  $\hat{\mu}_x^{**} = \bar{W}$  and  $\hat{\mu}_y^{**} = \bar{Y}$ ,

$$\bar{Y} = \frac{1}{rm} \sum_i \sum_k Y_{i[i]k}$$

where  $\bar{W} = \frac{1}{rm} \sum_i \sum_k W_{i(i)k}$ , and then the

estimate of population ratio R using DRSS is given by

$$\widehat{R}_{DRSS} = \frac{\bar{Y}}{\bar{W}} \tag{4.4}$$

By using Taylor expansion and assuming large population size, it is easy to show that

$$E(\widehat{R}_{DRSS}) = \frac{\mu_y}{\mu_x} + O\left(\frac{1}{n}\right),$$

and the variance of  $\widehat{R}_{DRSS}$  will be approximated by

$$\text{var}(\widehat{R}_{DRSS}) \cong \frac{R^2}{n} \left( V_x^2 + V_y^2 - 2\rho V_x V_y - m \left( \frac{\sum_{i=1}^r T_{w(i)}^{**2}}{n\mu_x^2} + \frac{\sum_{i=1}^r T_{y[i]}^{**2}}{n\mu_y^2} - 2 \frac{\sum_{i=1}^r T_{w(i)y[i]}^{**}}{n\mu_x\mu_y} \right) \right) \tag{4.5}$$

where  $V_x^2$  and  $V_y^2$  as in equation (4.2).

**Ratio Estimation Using DMRSS**

Using similar modification for bivariate case, and assuming that ranking is on variable X in the two stages. Then as in section 2,  $(W_{1(s)k}, Y_{1[s]k}), (W_{2(s)k}, Y_{2[s]k}), \dots, (W_{r(s)k}, Y_{r[s]k})$   $k=1, 2, \dots, m$  will denote the bivariate DMRSS where  $s$  is  $(\frac{r}{2})$  for the first  $\frac{r}{2}$  units and  $(\frac{r}{2} + 1)$  for the last  $\frac{r}{2}$  units in case when  $r$  is even and  $(\frac{r+1}{2})$  when  $r$  is odd.

$W_{i(s)k}$  is the  $s$ -th smallest X unit in the  $k$ -th cycle of the  $i$ -th bivariate MRSS in the first stage and  $Y_{i[s]k}$  is the corresponding Y observation in the  $k$ -th cycle of the  $i$ -th bivariate MRSS.

Two cases are considered here:

Case(1): When  $r$  is odd, the estimate of the population ratio R using  $DMRSS_o$  is defined by

$$\widehat{R}_{DMRSS_o} = \frac{\bar{Y}_o}{\bar{W}_o}, \quad \text{where } ,$$

$$\bar{W}_o = \frac{1}{mr} \sum_k \sum_i W_{i(\frac{r+1}{2})k}, \quad \bar{Y}_o = \frac{1}{mr} \sum_k \sum_i Y_{i[\frac{r+1}{2}]k} \tag{4.6}$$

Again, by using Taylor expansion we have

$$E(\widehat{R}_{DMRSS_o}) = \frac{\mu_y}{\mu_x} + O\left(\frac{1}{n}\right) \text{ and}$$

$$\text{var}(\widehat{R}_{DMRSS_o}) \cong \frac{R^2}{mr} (V_{w(s)}^{**2} + V_{y[s]}^{**2} - 2\rho_{w(s)y[s]} V_{w(s)}^{**} V_{y[s]}^{**}),$$

$$\text{where } s = \frac{r+1}{2} \tag{4.7}$$

$$V_{w(s)}^{**2} = \frac{\sigma_{w(s)}^{**2}}{\mu_x^2}, \text{ and } V_{y[s]}^{**2} = \frac{\sigma_{y[s]}^{**2}}{\mu_y^2}.$$

Case(2): When  $r$  is even, the estimate of the population ratio R using  $DMRSS_E$  is given by

$$\widehat{R}_{DMRSS_E} = \frac{\bar{Y}_E}{\bar{W}_E}, \tag{4.8}$$

$$\bar{W}_E = \frac{1}{mr} \sum_k \sum_{i=1}^r W_{i(s)k}$$

where,

$$= \frac{1}{mr} \sum_k \left( \sum_{i=1}^{r/2} W_{i(\frac{r}{2})k} + \sum_{i=1}^{r/2} W_{i(\frac{r}{2}+1)k} \right)$$

$$\bar{Y}_E = \frac{1}{mr} \sum_k \sum_{i=1}^r Y_{i[s]k}$$

and

$$= \frac{1}{mr} \sum_k \left( \sum_{i=1}^{r/2} Y_{i[\frac{r}{2}]k} + \sum_{i=1}^{r/2} Y_{i[\frac{r}{2}+1]k} \right).$$

Again by using Taylor expansion, and assuming symmetric underlying distributions then

$$E(\widehat{R}_{DMRSS_E}) = \frac{\mu_y}{\mu_x} + O\left(\frac{1}{n}\right) \text{ and}$$

$$\begin{aligned} \text{Var}(\widehat{R}_{\text{DMRSS}_E}) &\cong \frac{\mu_y^2}{\mu_x^2} \\ &\left( \frac{\sigma_{w(s)}^{**2}}{mr\mu_x^2} + \frac{\sigma_{y[s]}^{**2}}{mr\mu_y^2} - \frac{\sigma_{w(\frac{r}{2})y[\frac{r}{2}]}^{**} + \sigma_{w(\frac{r}{2}+1)y[\frac{r}{2}+1]}^{**}}{mr\mu_x\mu_y} \right) \\ &= \frac{R^2}{mr} \left( V_{w(s)}^{**2} + V_{y[s]}^{**2} - 2\rho_{w(s)y[s]} V_{w(s)}^{**} V_{y[s]}^{**} \right) \end{aligned} \tag{4.9}$$

where  $V_{w(s)}^{**2} = \frac{\sigma_{w(s)}^{**2}}{\mu_x^2}$ ,  $V_{y[s]}^{**2} = \frac{\sigma_{y[s]}^{**2}}{\mu_y^2}$  and s can

be either  $\frac{r}{2}$  or  $\left(\frac{r}{2} + 1\right)$  since

$$\sigma_{w(\frac{r}{2})}^{**} = \sigma_{w(\frac{r}{2}+1)}^{**} \text{ and } \sigma_{y[\frac{r}{2}]}^{**} = \sigma_{y[\frac{r}{2}+1]}^{**}.$$

**Ratio Estimation Using DERSS**

Assume without loss of generality that r is even. The case when r is odd is similar and it will be indicated in the numerical results only. Also assume ranking is on variable X. Let

$(W_{(1)k}, Y_{[1]k}), (W_{2(r)k}, Y_{2[r]k}), (W_{3(1)k}, Y_{3[1]k}), \dots, (W_{r(r)k}, Y_{r[r]k})$  be the bivariate

*DERSS* (see Samawi, 2001). This set of bivariate observations is independent but not identically distributed.

The estimate of the population ratio R using *DERSS* is given by

$$\widehat{R}_{\text{DERSS}} = \frac{\overline{Y}_{\text{DERSS}}}{\overline{W}_{\text{DERSS}}}, \tag{4.10}$$

where

$$\begin{aligned} \overline{W}_{\text{DERSS}} &= \frac{1}{mr} \sum_k \left( \sum_{i \text{ odd}} W_{i(1)k} + \sum_{i \text{ even}} W_{i(r)k} \right) \\ &= \frac{1}{mr} \sum_k \left( \sum_{i=1}^{r/2} W_{2i-1(1)k} + \sum_{i=1}^{r/2} W_{2i(r)k} \right) \\ \overline{Y}_{\text{DERSS}} &= \frac{\overline{Y}_{[1]} + \overline{Y}_{[r]}}{2}, \end{aligned}$$

$$\overline{Y}_{[1]} = \frac{1}{m} \sum_k \sum_{i=1}^{r/2} \left( \frac{Y_{2i-1[1]k}}{r/2} \right),$$

and  $\overline{Y}_{[r]} = \frac{1}{m} \sum_k \sum_{i=1}^{r/2} \left( \frac{Y_{2i[1]k}}{r/2} \right)$

Once again, by using Taylor expansion we have

$$E(\widehat{R}_{\text{DERSS}}) = \frac{\mu_y}{\mu_x} + O\left(\frac{1}{n}\right) \text{ and}$$

$$\begin{aligned} \text{Var}(\widehat{R}_{\text{DERSS}}) &\cong \frac{R^2}{mr} \left( V_{w(1)}^{**2} + V_{y[1]}^{**2} - 2\rho_{w(1)y[1]} V_{w(1)}^{**} V_{y[1]}^{**} \right) \\ &= \frac{R^2}{mr} \left( V_{w(r)}^{**2} + V_{y[r]}^{**2} - 2\rho_{w(r)y[r]} V_{w(r)}^{**} V_{y[r]}^{**} \right) \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} V_{w(j)}^{**} &= \sigma_{w(j)}^{**} / \mu_x, \quad V_{y[j]}^{**} = \sigma_{y[j]}^{**} / \mu_y \\ \text{and } \sigma_{w(j)y[j]}^{**} &= \rho_{w(j)y[j]} \sigma_{w(j)}^{**} \sigma_{y[j]}^{**}, \quad j=1 \end{aligned}$$

or r.

**Simulation Study**

A computer simulation is conducted to study the efficiency of estimating R when ranking is performed on the variable X. Using SRS, RSS, MRSS, ERSS, DRSS, DERSS and DMRSS, bivariate random samples were generated from a bivariate normal distribution with  $\mu_x=2$ ,  $\mu_y=4$ ,  $\sigma_x=1$ ,  $\sigma_y=1$  and  $\rho = \pm 0.9, \pm 0.8, \pm 0.5$ .

The performance of the ratio estimate will be investigated for r=4, 5, 6 and 7 and m=4 and 6. The ratios of the population means are estimated from SRS, RSS, MRSS, ERSS, DRSS, DERSS and DMRSS data sets. Using 5000 replications,

estimates of the means, the mean square errors and the ratio of the mean square errors (relative efficiency) for the ratio were computed.

Results of the simulation study

The values obtained by the simulation study are given in Table 4.1. In all cases the simulation showed that the efficiency of estimating R is not affected by the cycle size m, an explanation for this is that m is canceled in the numerator and dominator when relative efficiency is used. The values in the tables vary from a value of m to another because of the simulation variation. When the underlying distribution is  $N_2(2,4,1,1,\rho)$ , Table 4.1 shows that estimating the population ratio using DMRSS is more efficient

than using SRS, RSS, and MRSS. Also, using DRSS to estimate the population ratio is more efficient than using SRS and RSS, and using DERSS is more efficient than using SRS, RSS and ERSS.

Moreover, using the definition of relative efficiency, the double sampling schemes can be compared with each other. Our simulation indicates that, estimating the population ratio using DMRSS is more efficient than using DRSS and DERSS. Also, whenever  $|\rho|$  increases the efficiency increases in all cases. Note that negative values of  $\rho$  give higher efficiency than the positive values.

Table 4.1 Efficiency of the estimators of R when ranking on X and (X,Y) has  $N_2(4,2,1,1,\rho)$

M	r	DMRSS relative to			DRSS relative to		DERSS relative to		
		SRS	RSS	MRSS	SRS	RSS	SRS	RSS	ERSS
$\rho=0.9$									
4	4	4.15	1.95	1.67	2.79	1.31	2.19	1.03	1.15
	5	5.47	1.14	1.88	3.33	1.30	2.62	1.02	1.16
	6	5.77	2.27	1.89	3.61	1.42	2.77	1.09	1.38
	7	6.05	2.15	1.82	3.78	1.34	2.94	1.04	1.26
6	4	4.24	2.07	1.83	2.85	1.39	2.36	1.15	1.27
	5	4.87	2.14	1.81	3.02	1.33	2.48	1.09	1.20
	6	5.67	2.33	1.87	3.49	1.44	2.75	1.13	1.37
	7	5.91	2.17	1.85	3.72	1.36	2.76	1.01	1.24
$\rho=0.8$									
4	4	3.74	1.81	1.62	2.69	1.30	2.24	1.08	1.23
	5	4.28	1.88	1.66	2.99	1.31	2.37	1.04	1.14
	6	4.44	1.77	1.58	3.27	1.31	2.56	1.02	1.24
	7	4.41	1.79	1.61	3.31	1.35	2.54	1.03	1.20
6	4	3.55	1.71	1.53	2.52	1.21	2.14	1.03	1.21
	5	4.09	1.91	1.70	2.72	2.37	2.38	1.11	1.24
	6	4.10	1.91	1.65	2.83	1.32	2.30	1.07	1.19
	7	4.22	1.80	1.59	2.96	1.26	2.43	1.04	1.18
$\rho=0.5$									
4	4	3.11	1.69	1.52	2.37	1.29	2.04	1.11	1.19
	5	3.60	1.84	1.63	2.67	1.36	2.20	1.12	1.14
	6	3.90	1.80	1.59	2.88	1.33	2.43	1.12	1.30
	7	3.52	1.56	1.40	2.68	1.19	2.33	1.03	1.24

6	4	2.99	1.68	1.52	2.36	1.33	2.01	1.13	1.16
	5	3.45	1.67	1.51	2.53	1.23	2.15	1.04	1.10
	6	3.66	1.66	1.47	2.85	1.29	2.26	1.02	1.20
	7	3.46	1.65	1.46	2.63	1.26	2.29	1.09	1.22
					$\rho = -0.5$				
4	4	5.12	2.22	1.99	3.17	1.37	2.55	1.11	1.24
	5	6.08	2.40	2.01	3.51	1.38	2.91	1.15	1.34
	6	7.67	2.55	2.01	4.14	1.51	2.94	1.07	1.37
	7	7.23	2.23	1.96	4.22	1.41	3.23	1.08	1.33
6	4	4.52	2.08	1.78	2.96	1.36	2.39	1.10	1.27
	5	5.63	2.39	2.01	3.32	1.41	2.74	1.16	1.34
	6	6.93	2.44	2.00	4.17	1.88	3.11	1.09	1.40
	7	6.91	2.23	1.95	4.21	1.72	3.09	1.00	1.26
					$\rho = -0.8$				
4	4	6.73	2.73	2.37	3.83	1.56	2.89	1.17	1.35
	5	8.77	3.15	2.58	4.19	1.51	3.06	1.10	1.29
	6	10.31	3.52	2.72	4.97	1.70	3.41	1.17	1.42
	7	12.95	3.64	3.03	5.72	1.61	3.83	1.08	1.36
6	4	5.90	2.54	2.19	3.49	1.50	2.71	1.17	1.32
	5	9.33	3.12	2.62	4.24	1.42	3.25	1.09	1.27
	6	9.07	3.30	2.71	4.42	1.61	3.13	1.14	1.43
	7	11.84	3.60	2.79	5.62	1.71	3.72	1.13	1.39
					$\rho = -0.9$				
4	4	6.76	2.18	2.43	3.50	1.45	2.72	1.13	1.34
	5	11.03	3.84	3.05	4.54	1.58	3.52	1.23	1.36
	6	12.91	3.83	3.13	5.20	1.54	3.66	1.08	1.45
	7	17.19	4.66	3.56	6.36	1.72	3.99	1.08	1.42
6	4	6.84	2.90	2.46	3.56	1.51	2.78	1.18	1.40
	5	10.63	3.69	2.92	4.41	1.53	3.40	1.18	1.36
	6	12.82	4.04	3.19	5.17	1.63	3.43	1.08	1.41
	7	16.33	4.50	3.50	5.80	1.60	3.77	1.04	1.37

Application To Real Data Set And Conclusions

We illustrate the double ranked sample mean estimation procedure using a real data set which consists of the height (Y) and the diameter (X) at breast height of 399 trees. See Platt et al. (1988) for a detailed description of the data set. The summary statistics for the data are reported in Table 5.1. Note that the correlation coefficient is  $\rho=0.908$ .

Table 5.1. Summary Statistics of trees data.

Variable	Mean	Variance
Height (Y) in feet	52.36	325.14
Diameter (X) in cm	20.84	310.11

In this article, ranking is performed on the variable X exactly measured. However, in practice ranking is done before any actual quantification. Using a set size  $r = 3$  and the cycle size  $m = 3$ , we draw bivariate SRS, DRSS, and DMRSS of size 9,

however DERSS is the same as DRSS in this case. Table 5.2 contains all the above proposed estimators and their estimated variances using the drawn samples.

Table 5.2. Results from the drawn samples.

Sample	Naïve Estimator of the Diameter (X)	9(Estimated Variance)*	Ratio Estimator	9(Estimated Variance)*
SRS	13.57	168.60	2.50	1.036
DRSS	19.39	148.37	2.29	0.633
DMRSS	15.89	131.35	2.15	0.297

Table 5.2 confirms the simulation results. However, this example is just to illustrate the application using the proposed estimators.

Finally, the theoretical and simulation results showed that the population mean estimator using DMRSS is an unbiased estimator for the population mean whenever the underlying distribution is assumed to be symmetric. Also, it was shown theoretically that the variance of this estimator is less than the variance of the sample mean using MRSS (the first stage). Although using numerical simulation it was noticed that the sample mean based on DMRSS is more efficient than using other sampling methods (see Table 3.1) with respect to there variances.

Note there are difficulties in selecting the DMRSS because of the similarity of the subjects from the first stage. However, in practice this is not a problem because the number of units we rank in the second stage will not exceed 5.

In ratio estimation using the two stage sampling for different schemes, the estimator of the population ratio of two variables was introduced and the variance in each case was derived. Our numerical study indicated that the two stage sampling is more efficient than the first stage sampling considering the same sampling scheme with respect to their variances.

Comparing the two stage sampling schemes, namely DRSS, DMRSS and DERSS, superiority in efficiency depends on the distribution of the bivariate variable. However, DMRSS was more efficient than DRSS and DERSS when the underlying distribution is the bivariate normal. Moreover, those efficiencies depend on the set size  $r$  and the strength of the correlation between  $X$  and  $Y$ .

## References

- Al-Saleh, M. F., & Al-Kadiri, M. A. (2000). Double ranked set sampling. *Statistics and probability letters*, 48(2), 205-212.
- Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A first course in order statistics*. New York: John Wiley and Sons, Inc.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sampling survey methods and theory*. Vol. 2. John Wiley and Sons.
- Kaur, A., Patil, G. P., Sinha, A. K., & Tailie, C. (1995). Ranked set sampling: an annotated bibliography. *Environmental and Ecological Statistics*, 2, 25-54.
- McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked set. *Australian Journal of Agricultural Research*, 3, 385-390.
- Muttlak, H. A. (1997). Median ranked set sampling. *Journal of Applied Statistical Science*, 6(4), 245-255.
- Patil, G. P., Sinha, A. K., & Taillie, C. (1999). Ranked set sampling: A bibliography. *Environmental and Ecological Statistics*, 6, 91-98.
- Platt, W. J., Evans, G. W., & Rathbun, S.L. (1988). The population dynamics of a long-lived conifer. *The American Naturalist*, 131, 391-525.
- Samawi, H. M. (2001, in press). On double extreme ranked set sample with application to regression estimator.
- Samawi, H. M. Ahmed, M. S., & Abu Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrical Journal*, 38, (5) 577-586.
- Samawi, H. M., & Al-Sageer, O. A. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. *Biometrical Journal*, 43(3), 357-373.
- Samawi, H. M., & Muttlak, H. A. (1996). Estimation of ratio using ranked set sampling. *Biometrical Journal*, 38 (6), 753-764.
- Samawi, H. M., & Muttlak, H. A. (2001). On ratio estimation using median ranked set sampling. *Journal of Applied Statistical Science*, 10(2), 89-98.
- Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the stratified sampling by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1-31.



Yang, H.(1980). On the Variance of median and some other statistics. *Bulletin of Institute Mathematics Academic Simulation*, 10, 197-204.

Appendix

Theorem 1: Let X be a random variable with symmetric distribution function F(x) and mean μ, then

(a)  $g_{(\frac{r+1}{2})}(w)$  is symmetric about μ, if r is odd.

(b) Without loss of generality, assume that  $\mu = 0$ , then  $W_{(\frac{r}{2})} \stackrel{d}{=} -W_{(\frac{r+1}{2})}$  and

hence  $\frac{\mu_{\frac{r}{2}}^{**} + \mu_{\frac{r+1}{2}}^{**}}{2} = 0$  if r is even, where

$$\mu_{\frac{r}{2}}^{**} = E(W_{(\frac{r}{2})}) \text{ and } \mu_{\frac{r+1}{2}}^{**} = E(W_{(\frac{r+1}{2})}).$$

Proof:

(a) Because  $f_{(\frac{r+1}{2})}(x)$  is symmetric

about 0 (see Arnold, et al. 1992), then by using (3.5)

$g_{(\frac{r+1}{2})}(-w) = g_{(\frac{r+1}{2})}(w)$ . Therefore,  $g_{(\frac{r+1}{2})}(w)$  is symmetric about 0.

(b) Using the fact that in case of symmetry  $X_{(\frac{r}{2})} \stackrel{d}{=} -X_{(\frac{r+1}{2})}$ , when r is even, and from (3.7) and (3.9), then it is clear that

$$g_{(\frac{r+1}{2})}(-w) = g_{(\frac{r}{2})}(w) \quad \text{Therefore,}$$

$W_{(\frac{r}{2})} \stackrel{d}{=} -W_{(\frac{r+1}{2})}$ . Also note that

$$E\left(W_{(\frac{r}{2})}\right) = E\left(-W_{(\frac{r+1}{2})}\right) \Rightarrow \mu_{(\frac{r}{2})}^{**} = -\mu_{(\frac{r+1}{2})}^{**}$$

and hence  $\frac{\mu_{(\frac{r}{2})}^{**} + \mu_{(\frac{r+1}{2})}^{**}}{2} = 0$ .

Lemma 1: Let  $W_{1(\frac{r+1}{2})k}, W_{2(\frac{r+1}{2})k}, \dots, W_{r(\frac{r+1}{2})k}$ ,  $k=1,2,\dots, m$  be the  $DMRSS_O$  when r is odd, and

$$W_{1(\frac{r}{2})k}, W_{2(\frac{r}{2})k}, \dots, W_{\frac{r}{2}(\frac{r}{2})k}, W_{1(\frac{r+1}{2})k},$$

$$W_{2(\frac{r+1}{2})k}, \dots, W_{\frac{r}{2}(\frac{r+1}{2})k}, k=1, 2, \dots, m \text{ be the}$$

$DMRSS_E$  when r is even. If the c.d.f. F(x) is symmetric about its mean μ, then  $\overline{W}_{DMRSS_O}$  and  $\overline{W}_{DMRSS_E}$  are unbiased estimators for μ.

Proof: The proof is a consequence of Theorem 1.

Theorem 2: If the random variable X has a symmetric distribution function F(x) about μ, then

$$Var(\overline{W}_{DMRSS}) \leq Var(\overline{X}_{MRSS}) \leq \frac{\sigma^2}{n}, \text{ where}$$

$$\overline{X}_{MRSS} = \begin{cases} \frac{1}{n} \sum_{k=1}^m \sum_{i=1}^{r/2} (X_{i(\frac{r}{2})k} + X_{i(\frac{r+1}{2})k}), & \text{if r is even} \\ \frac{1}{n} \sum_{k=1}^m \sum_{i=1}^r X_{i(\frac{r+1}{2})k}, & \text{if r is odd} \end{cases}$$

Proof: Because Yang (1982) showed that  $Var(X_{(med)}) \leq \sigma^2$ ,

where  $X_{(med)}$  is the sample median of i.i.d sample of size r, then we need to prove only that  $Var(\overline{W}_{DMRSS}) \leq Var(\overline{X}_{MRSS})$ .

Case 1: When r is odd. Because  $W_{1(\frac{r+1}{2})k}, W_{2(\frac{r+1}{2})k},$

$\dots, W_{r(\frac{r+1}{2})k}, k=1,2,\dots,m$ , are i.i.d. from

$G_{(\frac{r+1}{2})}(w)$  then the prove is similar to that by

Yang (1982).

Case 2: When  $r$  is even,  $W_{i\binom{r}{2}k}$ ,  $i = 1, 2, \dots, \frac{r}{2}$ ,  $k=1, 2, \dots, m$  are i.i.d. with (3.8) distribution function and  $W_{1\binom{r+1}{2}k}, W_{2\binom{r+1}{2}k}, \dots, W_{\frac{r}{2}\binom{r+1}{2}k}$ , are i.i.d. with the distribution function as in (3.10). The two samples are independent. Also, assuming that  $\mu = 0$  we have that  $W_{\binom{r}{2}k} \stackrel{d}{=} W_{\binom{r+1}{2}k}$ , then by using Yang (1982)

$$\text{Var} \left( \frac{W_{\binom{r}{2}} + W_{\binom{r+1}{2}}}{2} \right) \leq \frac{\sigma_{\binom{r}{2}}^2 + \sigma_{\binom{r+1}{2}}^2}{2} = \sigma_{\binom{r}{2}}^2$$

, and hence  $\text{Var}(\bar{W}_{DMRSS_E}) \leq \text{Var}(\bar{X}_{MRSS_E})$ .

## On Distribution Function Estimation Using Double Ranked Set Samples With Application

Walid A. Abu-Dayyeh  
Department of Statistics  
Yarmouk University, Irbid Jordan

Hani M. Samawi  
Dept. of Mathematics & Statistics  
Sultan Qaboos University  
Al-Khod, Sultanate of Oman

Lara A. Bani-Hani  
Department of Statistics  
Yarmouk University, Irbid Jordan

---

As a variation of ranked set sampling (RSS); double ranked set sampling (DRSS) was introduced by Al-Saleh and Al-Kadiri (2000), and it has been used only for estimating the mean of the population. In this paper DRSS will be used for estimating the distribution function (cdf). The efficiency of the proposed estimators will be obtained when ranking is perfect. Some inference on the distribution function will be drawn based on Kolomgrov-Smirnov statistic. It will be shown that using DRSS will increase the efficiency in this case.

Key words: Double ranked set sample, distribution function estimation, Kolomgrov-Smirnov, ranked set.

---

### Introduction

In some practical situations, collecting units from the population is not too costly comparing with quantification of the sampling units. A large number of those units may be identified to represent the population of interest and yet only a carefully selected subsample is to be quantified. This potential for observational economy was recognized for estimating the mean pasture and forge by McIntyre (1952). He proposed a method, later called ranked set sampling (RSS) by Halls and Dell (1966), currently under active investigation.

RSS procedure can be described as follows: Identify a group of sampling units randomly from the target population. Then, randomly partition the group into disjoint subsets each having a pre-assigned sizer  $r$ , in the most practical situations, the size  $r$  will be 2, 3 or 4. Then, rank each subset by a suitable method of ranking such as prior information, visual inspection or by the experimenter himself.

In terms of sampling notation,

where  $X_{j(i)}$  denotes the  $i$ -th ordered statistic in the  $j$ -th set. Then the  $i$ -th ordered statistic from the  $i$ -th subset will be quantified,  $i = 1, \dots, r$ . Then  $X_{1(1)}, X_{2(2)}, \dots, X_{r(r)}$  will be obtained. The whole process can be repeated  $k$ -times, to get a RSS of size  $n = kr$ . The resulting sample is called the balanced ranked set sample (RSS). Through all the paper, only balanced RSS will be used.

Al-Saleh and Al-Kadiri (2000) extended RSS to double rank set sample (DRSS). DRSS can be described as follows:

1. Identify  $r^3$  elements from the target population and divide these elements randomly into  $r$  subsets each of size  $r^2$  elements.
2. Use usual RSS procedure to obtain  $r$  RSS each of size  $r$ .
3. Apply again the RSS procedure in Step 2, on the  $r$  RSS's.

We may repeat steps 1, 2 and 3  $k$ -times to obtain DRSS sample of size  $n = rk$ . In DRSS, ranking in the second stage is easier than ranking in the first stage, (see Al-Saleh and Al-Kadiri, 2000).

Moreover, an up-to-date annotated bibliography for RSS can be found in Kaur et al., (1995) and Patil et al. (1999). Stokes and Sager (1988) estimate the distribution functions,  $F(x)$  say, for a random variable  $X$  by the empirical cdf ( $F^*$ ) based on the RSS, which will be given in

---

The contact person for this article is Hani M. Samawi. Email him at [hsamawi@squ.edu.om](mailto:hsamawi@squ.edu.om).

Section 2. They pointed out that,  $F^*(t)$  is an unbiased for  $F(t)$  and is more efficient than the empirical distribution function of a SRS

$(\hat{F}(t))$  of size  $n$  with

$$\text{Var}(F^*(t)) = \frac{1}{kr^2} \sum_{i=1}^m F_i(t) [1 - F_i(t)], \quad \text{where}$$

$$F_i(t) = F_{(i)}(t) = I_{F(t)}(i, r - i + 1) \quad (1.1)$$

for perfect ranking, and  $I_{F(t)}(i, r - i + 1)$  is the incomplete beta ratio function.

Basic Setting of DRSS

Let  $Y_1, \dots, Y_r$  be a DRSS, and assume that  $Y_i \sim g_i(y)$  with df, mean and variance are:  $G_i(y), \mu_i^*$  and  $\sigma_i^{*2}$ , respectively. Al-Saleh and Al-Kadiri (2000) showed that:

(i)  $f(y) = \frac{1}{r} \sum_{i=1}^r g_i(y), \quad (1.2)$

(ii)  $F(y) = \frac{1}{r} \sum_{i=1}^r G_i(y), \quad (1.3)$

(iii)  $\mu = \frac{1}{r} \sum_{i=1}^r \mu_i^*, \quad (1.4)$

(iv)

$$\sigma^2 = \frac{1}{r} \left[ \sum_{i=1}^r \sigma_i^{*2} + \sum_{i=1}^r (\mu_i^* - \mu)^2 \right], \quad (1.5)$$

where  $f, F, \mu$  and  $\sigma^2$  are the pdf, cdf, mean and variance of the population.

In this paper, we will consider the problem of estimating the distribution function  $F$  using DRSS. In Section 2, the empirical cdf estimator based on DRSS ( $\hat{FDR}$ ) will be considered. The efficiency between the DRSS estimator and those estimators based on SRS and RSS will be obtained when ranking is perfect. In Section 3 the Kolmogrov-Smirnov statistic will be studied based on a DRSS. Also, a confidence interval of  $F(t)$  will be constructed using the Kolmogrov-Smirnov statistic based on DRSS.

Estimating The Distribution Functions Using DRSS

In this Section the distribution function will be estimated using the DRSS, in the cases where ranking is perfect and when ranking is imperfect. The suggested estimator will be compared with the cdf estimators based on SRS and RSS via their variances.

Definition and Some Basic Results

For the  $l$ -th cycle, let  $\{Y_{1l}, Y_{2l}, \dots, Y_{rl}\}, l = 1, \dots, k$ , be a DRSS of size  $r$ , and assume that  $Y_i$  has the probability density function (pdf)  $g_i(y)$  and the cdf  $G_i(y)$ . Note that  $g_i(y)$  is the density of the  $i$ -th ordered statistic of a RSS with densities  $f_{(1)}, f_{(2)}, \dots, f_{(r)}$  and distribution functions  $F_{(1)}, F_{(2)}, \dots, F_{(r)}$  respectively. Then

$$G_i(y) = \sum_{j=i}^r \sum_{S_j} \prod_{L=1}^j F_{(L)}(t) \prod_{L=j+1}^r [1 - F_{(L)}(t)] \quad (2.1)$$

where the set  $S_i$  consists of all permutations  $(i_1, i_2, \dots, i_r)$  of  $1, 2, \dots, r$  for which  $i_1 < \dots < i_j$  and  $i_{j+1} < \dots < i_r$  (see Al-Saleh and Al-Kadiri, 2000).

Let  $\hat{FDR}, \hat{F}$  and  $F^*$  be the edf's (empirical distribution functions) of DRSS, SRS and RSS from the population with cdf  $F$ , then:

$$\hat{FDR}(t) = \frac{1}{kr} \sum_{j=1}^k \sum_{i=1}^r I[Y_{ij} \leq t] \quad (2.2)$$

$$\hat{F}(t) = \frac{1}{kr} \sum_{i=1}^{rk} I[X_i \leq t] \quad (2.3)$$

$$F^*(t) = \frac{1}{kr} \sum_{j=1}^k \sum_{i=1}^r I[X_{i(i)} \leq t] \quad (2.4)$$

respectively, where  $I(\cdot)$  is the indicator function. Then, we have the following results.

a)  $E[\hat{FDR}(t)] = F(t)$

$$b) \text{ var}(\hat{FDR}(t)) = \frac{1}{kr^2} \sum_{i=1}^r G_i(t)[1 - G_i(t)], \tag{2.5}$$

(see the Appendix for the prove of these results.) Also, we show in the Appendix that

$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}(\hat{FDR}(t))]^{1/2}$  converges in distribution to a standard normal random variable as  $k \rightarrow \infty$  when  $r$  and  $t$  are held fixed. Moreover, it can be shown that an unbiased estimator of

$\text{var}[\hat{FDR}(t)]$  is given by

$$\hat{\text{var}}[\hat{FDR}(t)] = \frac{1}{(k-1)r^2} \sum_{i=1}^r \hat{G}_i(t)[1 - \hat{G}_i(t)], \tag{2.6}$$

where  $\hat{G}_i(t) = \frac{1}{k} \sum_{j=1}^k I[Y_{ij} \leq t]$  is the edf based

on all  $k$  of the  $i$ -th judgment order statistic and hence it can be shown also that

$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\hat{\text{var}}[\hat{FDR}(t)]]^{1/2}$  converges in distribution to a standard normal random variable as  $k \rightarrow \infty$  when  $r$  and  $t$  are held fixed. (See the Appendix for the prove of the above results.) Therefore, when  $k$  is large for a specified value  $t$ , an approximate  $100(1-\alpha)\%$  confidence interval for  $F(t)$  is

$$\hat{FDR}(t) \pm Z_{\alpha/2} \sqrt{\hat{\text{var}}[\hat{FDR}(t)]} \tag{2.7}$$

Finally, as a special case when  $r = 2$ , it can be shown that  $\text{var}[\hat{FDR}(t)] \leq \text{var}[\hat{F}(t)]$  and  $\text{var}[\hat{FDR}(t)] \leq \text{var}[F^*(t)]$ . (See the Appendix Lemma 2 for the prove of this results.)

### Efficiency of $\hat{FDR}$

The edf is used for making pointwise estimates of  $F(t)$ , as well as for making inference concerning the overall population distribution. In this section, we will examine the magnitude of the

improvement in precision that results when estimating  $F(t)$  by  $\hat{FDR}(t)$  rather than by  $\hat{F}(t)$  or  $F^*(t)$ .

Now, the relative precision (RP) of the double ranked set to the simple random sampling estimator and to ranked set sample estimator, are defined by:

$$RP_1(t) = \frac{\text{var}[\hat{F}(t)]}{\text{var}[\hat{FDR}(t)]} = \frac{F(t)[1 - F(t)]}{F(t) - \left[ \frac{\sum_{i=1}^r G_i^2(t)}{r} \right]} \tag{2.8}$$

$$RP_2(t) = \frac{\text{var}[F^*(t)]}{\text{var}[\hat{FDR}(t)]} = \frac{rF(t) - \sum_{i=1}^r F_{(i)}^2(t)}{rF(t) - \sum_{i=1}^r G_{(i)}^2(t)} \tag{2.9}$$

Table 1 and Table 2 show the value of  $RP_1(F^{-1}(p))$  and  $RP_2(F^{-1}(p))$  respectively, for some values of  $p$  and  $r = 2, 3, 4, 5$ . It can be noticed that both of  $RP_1$  and  $RP_2$  are monotone increasing from  $p = 0$  to  $p = 0.5$ , to achieve their maximum at  $p = 0.5$ . Also, they are symmetric about  $p = 0.5$ . Table 1 and Table 2 show that the gain in efficiency from DRSS for estimation of  $F(t)$  is substantial when the ranking can be done perfectly.

Table 1.  $RP_1(F^{-1}(p))$  when ranking of X is perfect.

		P							
r	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5	
	1	5	0	5	0	0	0	0	
2	1.0	1.0	1.1	1.1	1.2	1.4	1.5	1.6	
	1	5	2	9	7	4	8	4	
3	1.0	1.1	1.2	1.4	1.6	1.9	2.0	2.1	
	2	1	6	2	0	1	8	2	
4	1.0	1.1	1.4	1.6	1.9	2.3	2.5	2.6	
	3	8	1	8	4	2	2	0	
5	1.0	1.2	1.5	1.9	2.2	2.8	3.4	4.2	
	4	5	8	5	9	8	3	7	

Table 2.  $RP_2(F^{-1}(p))$  when ranking of X is perfect.

		P							
R	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5	
	1	5	0	5	0	0	0	0	
2	1.0	1.0	1.0	1.0	1.0	1.1	1.2	1.2	
	0	0	2	4	7	4	0	3	
3	1.0	1.0	1.0	1.1	1.1	1.2	1.3	1.3	
	0	1	5	1	7	8	2	3	
4	1.0	1.0	1.1	1.1	1.2	1.3	1.4	1.4	
	0	3	0	8	6	6	0	2	
5	1.0	1.0	1.1	1.2	1.3	1.5	1.7	2.1	
	0	4	4	6	6	3	2	0	

Inference on the distribution function

Because the distribution function F can be estimated more efficiently from a double ranked set sample than from a SRS and a RSS, it suffices to note that the statistics based on an estimate of F(t), such as the Kolmogorov-Smirnov statistic, would be improved in some sense as well.

In particular, we observe that the null distribution of the statistic

$D^{**} = \sup_t [\hat{FDR}(t) - F_0(t)]$  is stochastically smaller than  $D^* = \sup_t [F^*(t) - F_0(t)]$  and smaller

than  $D = \sup_t [\hat{F}(t) - F_0(t)]$  when  $D^{**}$ ,  $D^*$  and  $D$  are all based on the same number of measured observations. We mean that

$$H_{(r)k}^{**}(d) \geq H_{(r)k}^*(d) \text{ and } H_{(r)k}^{**}(d) \geq H_{(r)k}(d)$$

with strict inequality for some d, where

$$H_{(r)k}^{**}(d) = p(D^* \leq d]$$

$$H_{(r)k}^*(d) = p(D^* \leq d) \text{ and } H_{rk}(d) = p(D \leq d].$$

Where D,  $D^*$ , and  $D^{**}$  are calculated from a SRS, a RSS and a DRSS of size rk respectively.

This implies that  $100(1-\alpha)\%$  of  $D^{**}$ ,

which be denoted by  $C_{\alpha}^{**}$ , will always be less

than or equal to corresponding percentile of the

statistics D and  $D^*$ , denoted by  $C_{\alpha}$  and  $C_{\alpha}^*$

respectively. A confidence band for F based on

$D^{**}$  is

$$\hat{FDR} \pm C_{\alpha}^{**}, \tag{3.1}$$

is narrower than the corresponding band based on D and  $D^*$ .

In this section, the simulations which we done, is true for some finite values of r and k in the case of perfect judgment ranking. To find the

table of critical values of  $D^{**}$  ( $C_{\alpha}^{**}$ ) we draw a

double ranked set sampling ( $Y_i$ 's) of size n from

uniform distribution with parameters 0, 1. Then all

elements in the sample will be rank ( $X_{(i)}$ 's).

Now for  $k=1$ ,

$$D^{**} = \max \left\{ \max_{1 \leq i \leq n} \left| \frac{1}{n} - F_0(Y_{(i)}) \right|, \max_{1 \leq i \leq n} \left| F_0(Y_{(i)}) - \frac{i-1}{n} \right|, 0 \right\}$$

where

$$F_0(X_{(i)}) = X_{(i)}.$$

The previous procedure will be repeated until we get,  $D_1^{**}, D_2^{**}, \dots, D_{10000}^{**}$ . Also,  $D_i^{**}$ 's will be ranked to find  $C_\alpha^{**}$  such that,  $P(D^{**} \leq C_\alpha^{**}) = 1 - \alpha$ , i.e., the  $C_\alpha^{**} = D_{(i)}^{**}$  where  $i = [(1 - \alpha)10000]$ , where  $[d]$  is the greatest interge of  $d$ .

Now, Table 3 reports the critical values  $C_\alpha^{**}$  for the test statistic  $D^{**}$  for  $\alpha = 0.01, 0.05$  and  $0.10$  for  $r = 2, 3, 4, 5$  and  $k = 2, 3, \dots, 20$ . The table shows that DRSS can result in a substantial decrease in width of the simultaneous confidence band. The amount of the improvement can be described by the quantities,

$$R_{\alpha 1} = \left( \frac{C_\alpha}{C_\alpha^{**}} \right)^2 \quad (3.2)$$

$$R_{\alpha 2} = \left( \frac{C_\alpha^*}{C_\alpha^{**}} \right)^2 \quad (3.3)$$

Because  $R_{\alpha 1}$  and  $R_{\alpha 2}$  are the square of the ratio of confidence-band widths, then they can be interpreted as a measure of relative precision. The ratios  $R_{\alpha 1}$  and  $R_{\alpha 2}$  are computed from the entries of Table 3 ( $C_\alpha^{**}$ ), Table 2 ( $C_\alpha^*$ ) (from Stokes and Sager; 1988) and the Table of critical values for the Kolmogrove-Smirnov statistic  $D$  (from Gibbons and Chakraborti (1992)).

Table 4 gives the values of  $R_{\alpha 1}$  and  $R_{\alpha 2}$  at  $r = 2, \dots, 5$  and  $k = 2, \dots, 10$ . These values are comparable with those of Table 1 and Table 2. So,  $R_{\alpha 1}$  and  $R_{\alpha 2}$  indicate the same thing which given by  $R_{p1}(t)$  and  $R_{p2}(t)$ , when ranking of  $X$  is perfect.

Table 3. Critical values of  $D^{**}(C_{\alpha}^{**})$ 

	r=2			r=3			r=4			r=5		
	$\alpha:$											
k	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
2	0.43	0.47	0.01	0.36	0.40	0.47	0.13	0.35	0.42	0.28	0.32	0.38
3	0.33	0.36	0.57	0.27	0.30	0.36	0.24	0.26	0.31	0.21	0.24	0.28
4	0.27	0.29	0.44	0.22	0.24	0.28	0.19	0.21	0.26	0.17	0.19	0.23
5	0.23	0.25	0.34	0.19	0.21	0.24	0.17	0.18	0.21	0.15	0.16	0.19
6	0.20	0.22	0.29	0.16	0.18	0.21	0.14	0.16	0.18	0.13	0.14	0.17
7	0.18	0.19	0.25	0.15	0.16	0.19	0.13	0.14	0.16	0.12	0.13	0.15
8	0.16	0.17	0.22	0.13	0.15	0.17	0.11	0.13	0.15	0.10	0.11	0.13
9	0.15	0.16	0.20	0.12	0.13	0.15	0.11	0.12	0.13	0.10	0.10	0.12
10	0.14	0.15	0.19	0.11	0.12	0.14	0.10	0.11	0.12	0.09	0.10	0.11
11	0.13	0.14	0.17	0.10	0.11	0.13	0.09	0.10	0.12	0.08	0.09	0.10
12	0.12	0.13	0.16	0.10	0.11	0.12	0.08	0.09	0.11	0.08	0.08	0.10
13	0.11	0.12	0.15	0.09	0.10	0.12	0.08	0.09	0.10	0.07	0.08	0.09
14	0.10	0.11	0.14	0.09	0.09	0.11	0.08	0.08	0.09	0.07	0.07	0.08
15	0.09	0.11	0.13	0.08	0.09	0.10	0.07	0.08	0.09	0.06	0.07	0.08
16	0.09	0.10	0.12	0.08	0.08	0.10	0.07	0.07	0.09	0.06	0.07	0.08
17	0.09	0.10	0.12	0.07	0.08	0.09	0.07	0.07	0.08	0.06	0.06	0.07
18	0.09	0.09	0.11	0.07	0.08	0.09	0.06	0.07	0.08	0.06	0.06	0.07
19	0.08	0.09	0.11	0.07	0.07	0.08	0.06	0.06	0.07	0.05	0.06	0.07
20	0.08	0.09	0.10	0.07	0.07	0.08	0.06	0.06	0.07	0.05	0.06	0.06



Table 4. The values of  $R_{\alpha 1}$  and  $R_{\alpha 2}$ 

$R_{\alpha 1}$									
	r= 2			r=3			r=4		
	$\alpha:$								
k	0.01	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
2	1.17	1.76	1.79	1.70	1.72	1.74	1.75	1.65	1.65
3	2.03	2.09	2.15	2.09	2.05	20.1	2.01	2.14	2.11
4	2.31	2.41	2.52	2.39	2.51	2.58	2.49	2.47	2.25
5	0.59	2.69	2.85	2.49	2.62	2.78	2.52	2.60	2.78
6	2.89	2.98	3.24	3.06	2.97	3.10	2.94	2.85	3.16
7	2.97	3.39	3.64	.300	3.29	3.20	3.13	3.19	3.52
8	3.52	3.77	3.80	3.41	3.24	3.45	3.64	3.13	3.48
9	3.48	3.75	3.79	3.67	3.70	4.27	3.30	3.36	4.31
10	3.72	3.74	4.24	4.00	4.00	4.29	4.00	3.64	4.34
$R_{\alpha 2}$									
2	1.41	1.42	1.34	1.23	1.16	1.18	1.15	1.27	1.22
3	1.70	1.70	1.62	1.49	1.44	1.43	1.36	1.33	1.27
4	1.88	2.00	2.08	1.74	1.78	1.74	1.60	1.54	1.42
5	2.19	2.19	2.30	1.87	1.78	20.1	1.67	1.78	1.78
6	2.40	2.39	2.56	2.25	2.09	2.18	2.04	1.72	2.09
7	2.60	2.66	2.83	2.15	2.25	2.33	1.92	2.04	2.07
8	2.85	3.11	3.06	2.61	2.35	2.52	2.39	2.14	2.15
9	2.78	3.06	3.02	2.78	2.86	2.78	2.12	2.25	2.61
10	2.94	3.00	.354	2.98	2.51	2.94	2.56	2.39	2.78

References

Al-Saleh, & Al-Kadiri (2000). Double ranked set sampling. *Statistics and Probability Letters*, 48(2), 205-212.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference*. Marcel Dekker, Inc. New York, Basel, Hong Kong.

Halls, L. K., & Dell, T. R. (1966). Trial of ranked set sampling for forage yields. *Forest Science*, 12, 22-26.

Kaur, A. Patil, G. P., Sinha, A. K. and Tailie, C. (1995). Ranked set sampling. An annotated bibliography. *Environmental and Ecological Statistics*, 2, 25-45.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked set. *Australian Journal of Agricultural Research*, 3, 385-390.

Patil G. P., A. K. Sinha and Tillie C. (1999). Ranked set sampling: A Bibliography. *Environmental Ecological Statistics*, 6, 91-98.

Stokes, S. L., & Sager, T. (1988). Characterization of a ranked set sample with application to estimating distribution functions. *Journal of American Statistical Association*, 83, 374-381.

Appendix

Proposition 1.  $\hat{FDR}$  is an unbiased estimator of F.

- a)  $E[\hat{FDR}(t)] = F(t)$
- b)

$$\text{var}(\hat{FDR}(t)) = \frac{1}{kr^2} \sum_{i=1}^r G_i(t)[1 - G_i(t)].$$

Proof:

From the definition of a DRSS the proof will follow simply by using (1.3) and (2.1).

Proposition 2.

$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}(\hat{FDR}(t))]^{1/2}$  converges in distribution to a standard normal random variable as  $k \rightarrow \infty$  when r and t are held fixed.

Proof: This follows from rewriting  $\hat{FDR}$  as

$$\hat{FDR} = \frac{1}{k} \sum_{j=1}^k U_j, \quad \text{where}$$

$$U_j = \sum_{i=1}^r \frac{I[Y_{ij} \leq t]}{r}, \text{ then } U_j \text{'s are iid,}$$

therefore the proof follows directly from the Central Limit Theorem.

Lemma 1.

- (a)  $\text{var}[\hat{FDR}(t)]$  is an unbiased estimator of

$$\text{var}[\hat{FDR}(t)].$$

where:

$$\text{var}[\hat{FDR}(t)] = \frac{1}{(k-1)r^2} \sum_{i=1}^r \hat{G}_i(t)[1 - \hat{G}_i(t)]$$

$$\text{and } \hat{G}_i(t) = \frac{1}{k} \sum_{j=1}^k I[Y_{ij} \leq t] \text{ is the edf}$$

based on all k of the i-th judgment order statistic.

- (b)

$$[\hat{FDR}(t) - E(\hat{FDR}(t))]/[\text{var}[\hat{FDR}(t)]]^{1/2}$$

converges in distribution to a standard normal random variable as  $k \rightarrow \infty$  when r and t are held fixed.

Proof:

- (a)

$$E[\text{var}[\hat{FDR}(t)]] = \frac{1}{(k-1)r^2} \sum_{i=1}^r [E(\hat{G}_i(t)) - E(\hat{G}_i^2(t))]$$

$$\text{because } E(\hat{G}_i(t)) = G_i(t)$$

and

$$E(\hat{G}_i^2(t)) = \frac{1}{k^2} \sum_{j=1}^k \text{var}(I[Y_{ij} \leq t]) + [G_i(t)]^2$$

$$\begin{aligned}
 &= \frac{k}{k^2} \text{var}(I[Y_i \leq t]) + G_i^2(t) \\
 &= \frac{G_i(t)[1 - G_i(t)]}{k} + \frac{kG_i^2(t)}{k} \\
 &= \frac{G_i(t) + (k - 1)G_i^2(t)}{k} .
 \end{aligned}$$

Then  $E[\text{var}(\hat{FDR}(t))] =$

$$\begin{aligned}
 &\frac{1}{(k-1)^2} \sum_{i=1}^r \left[ \frac{kG_i(t)}{k} - \frac{G_i(t) + (k-1)G_i^2(t)}{k} \right] \\
 &= \frac{1}{kr^2} \sum_{i=1}^r G_i(t)[1 - G_i(t)] \\
 &= \text{var}[\hat{FDR}(t)]
 \end{aligned}$$

Part (b) can be shown by noting that:

$$\frac{\text{var}(\hat{FDR}(t))}{\text{var}(FDR(t))} \xrightarrow{p \rightarrow 1} 1 \text{ as } k \rightarrow \infty, \text{ and}$$

because  $\hat{G}_i(t) \xrightarrow{p} G_i(t)$ .

Furthermore, by Lemma 1 when  $k$  is large for a specified value  $t$ , an approximate  $100(1-\alpha)\%$  confidence interval for  $F(t)$  is:

$$\hat{FDR}(t) \pm Z_{\alpha/2} \sqrt{\widehat{\text{var}}[\hat{FDR}(t)]}$$

Lemma 2. : For the special case when  $r = 2$ ,

(a)  $\text{var}[\hat{FDR}(t)] \leq \text{var}[\hat{F}(t)]$

(b)  $\text{var}[\hat{FDR}(t)] \leq \text{var}[F^*(t)]$ .

Proof: Let  $k = 1$  and  $F(t) = F$

$$F_1(t) = 2F - F^2,$$

$$F_2(t) = F^2, \quad G_1(t) = F^4 - 2F^3 + 2F, \text{ and}$$

$$G_2(t) = 2F^3 - F^4.$$

Then  $\text{var}[\hat{F}(t)] = \frac{2F - 2F^2}{4}$

$$\text{var}[\hat{F}^*(t)] = \frac{-2F^4 + 4F^3 - 4F^2 + 2F}{4},$$

and  $\text{var}[\hat{FDR}(t)] = \frac{1}{4}[-2F^8 + 8F^7 - 8F^6 - 4F^5 + 8F^4 - 4F^2 + 2F]$ .

Then  $\text{var}[\hat{FDR}(t)] = \text{var}[\hat{F}(t)] - \frac{2F^2(1-F)^2(F^4 - 2F^3 - F^2 + 2F + 1)}{4} \leq \text{var}[\hat{F}(t)]$

Also,  $\text{var}[\hat{FDR}(t)] = \text{var}[F^*(t)] - \frac{2F^3[1-F]^3[2-F][F+1]}{4} \leq \text{var}[F^*(t)]$ ,

$0 \leq F \leq 1$ .

## Type I Error Rates For Rank-Based Tests Of Homogeneity Of Slopes

Alan J. Klockars  
Educational Psychology  
University of Washington

Tim P. Moses  
Educational Psychology  
University of Washington

---

The purpose of this study was to explicate two issues concerning the standard and rank based test of homogeneity of slopes. Two alternative ranking methods intended to address nonnormality and additive treatment effect patterns were developed and compared in terms of their ability to control Type I error. The results replicated previous findings of inflated Type I error rates with leptokurtic curves and with rank based tests with some patterns of additive treatment effects. The new nonparametric procedures generally control Type I error although they were slightly inflated with skewed distributions.

Key words: Slope homogeneity, ranking methodology, type I error

---

### Introduction

Psychology and education have long acknowledged the need for methods to address the interaction between treatment variables on the one hand and individual difference variables on the other. Cronbach (1957) in his presidential address to the American Psychological Association called for a fusion of the “two schools of psychology”, a field later to be identified as Aptitude x Trait interaction (ATI) research (Cronbach & Snow, 1981). While ATI research was originally developed within educational psychology it has spread throughout psychology including industrial psychology (see for instance, Hunter, Schmitt & Hunter, 1979) and psychotherapy (see for instance

Dance & Neufeld, 1988). Two major strategies are used to explore ATIs. The first is based on stratification of the individual difference variable, which produces a randomized block design. The desired information is contained in the Block x Treatment interaction. The alternative is a regression based approach that can be viewed either as a test of moderated regression or of homogeneity of slopes within an analysis of covariance design.

The usual form of the regression approach is to assume a linear relationship between the individual difference variable used as the covariate (X) and the outcome measure (Y). The issue investigated is whether the treatment alters the nature of the linear relationship. The presence of an interaction between the treatment and X is reflected in the difference between the slopes. This finding may be the primary finding of the study and may also inform the researcher regarding appropriate strategies for looking for main effects.

We will adopt the regression vantage point for describing the issues addressed. Cronbach and Snow (1981) argued for the regression approach as more powerful than stratification, an assertion that was supported in simulations by Klockars and Beretvas (2001). The issue of power is particularly important given the high Type II error rates associated with attempts to identify interactions, especially in field studies (McClelland & Judd, 1993). For a comparison of randomized block and analysis of covariance see

---

Alan J. Klockars is Professor of Educational Psychology at the University of Washington. His research concerns multiple comparisons and, more recently, methods of conducting ATI research. Address correspondence to Alan J. Klockars, Box 353600, University of Washington, Seattle, WA 98195-3600. E-mail: klockars@u.washington.edu. Tim P. Moses is a doctoral candidate in the Educational Psychology program at the University of Washington. His research and teaching focus on the application of statistics to the study of social phenomena and the influences of assumption violations on the accuracy of standard and alternative statistical methods.

Klockars, Potter, and Beretvas (1999), and Klockars and Beretvas (2001).

The test of homogeneity of slopes is based on a set of assumptions common to both regression and covariance. Of primary importance in the current investigation is the assumption that the variables are normally distributed. The assumption is part of a mathematical model and, as with any model, it is unexpected that empirical data will ever exactly fulfill the model (e.g. scores are discrete while the model is continuous).

However, Micceri (1989) in a survey of typical variables analyzed in psychology and education journals reported that the distributions were often far from normal with considerable skew and kurtosis. Conover and Iman (1982) and more recent work by Headrick and Sawilowsky (2000) showed that the Type I error control of the test of homogeneity of slopes is greatly impacted by the shape of the distributions involved. Platykurtic or light-tailed distributions produce Type I error rates that are conservative while leptokurtic or heavy-tailed distributions produce liberal Type I error rates. Klockars and Moses (2001) found that the Type I error rates for distributions with shapes that Micceri (1989) indicated were typical far exceeded both Bradley's (1978) conservative (.055) and liberal (.075) definition of robustness.

Prior research has not directly addressed the question of the relative impact of nonnormality in X and Y on Type I error. Atiquallah (1964) showed analytically that the shape of the distribution of X plays a role in the magnitude of the calculated F ratio as does the distribution of Y. In simulation studies three different patterns of X and Y distributions have been used.

Conover and Iman (1982) and Stephenson and Jacobson (1988) varied the shape of the Y distribution but used a normally distributed X distribution throughout. Headrick and Sawilowsky (2000) let the X and Y distributions have the same shape so that if Y were moderately right skewed the X distribution would also be moderately right skewed. Klockars and Moses (2001) systematically varied the shape of the Y distribution and created the X distribution as a linear combination of Y and normally distributed random error. Thus the covariate, X, had a distribution less extreme than that of the Y distribution. This was particularly true with the

low correlation condition in which the normally distributed random error was more heavily weighted.

The first issues under investigation in the current study are (1) a replication of the finding that the shape of the Y distribution systematically influences Type I error rates of the test of homogeneity of slopes, and (2) an evaluation of the relative importance and independence of the shape of the X distribution compared to that of the Y distributions in producing Type I errors.

A number of authors have proposed non-parametric, rank based analyses of covariance to avoid the distributional requirements of analysis of covariance as a test of adjusted means (Quade, 1967; Puri & Sen, 1969; Burnett & Barr, 1977; Shirley, 1981). These strategies, however, focused primarily on the null hypothesis regarding the adjusted means of the treatment groups. Slopes were assumed to be homogeneous and the question of an interaction was not addressed.

Shirley (1981) developed  $\chi^2$  tests for both the test of parallel lines and equal adjusted means on data where the outcome measure Y was converted to ranks. Conover and Iman (1982) proposed standard analysis of covariance on data where both X and Y were replaced with their ranks. Stevenson and Jacobsen (1988) offered a "hybrid" alternative in which only the Y variable was ranked while X was retained in its original form. A standard ANCOVA was conducted on the raw X and ranked Y scores to test for both differences in slopes and adjusted means. In the latter two studies simulated data were generated to evaluate how robust the methods were. The rank and hybrid ANCOVA methods tended to control Type I error in situations where the error rate for the original observations was problematic, that is, where the Y distributions were leptokurtic.

More recent inquiries using analysis of covariance with ranks have returned to considering only questions about the adjusted means (Seaman, Algina, & Olejnik, 1985; Harwell & Serlin, 1988; Hettermansperger, 1984; Rheinheimer & Penfield, 2001). However, Headrick and Sawilowsky (2000) presented simulation evidence that indicated that the Conover and Iman approach to testing differences in slopes can have very elevated Type I error rates under conditions of additive treatment effects. In particular, simulations in which a small proportion of the treatment effects had large

additive effects resulted in extremely high Type I error rates when the test for homogeneity of slopes was conducted. When X and Y were highly correlated and the sample size was large there was essentially a 100% chance of rejecting the null hypothesis that the slopes differed. This happened even though the only effects built into the data were additive effects that should have been reflected in the test of adjusted means rather than slopes. The present study (3) replicates the Headrick and Sawilowsky finding and (4) develops alternative methods for testing for differences in slopes within the general analysis of covariance framework that may have better control of Type I error.

The development of alternative non-parametric methods relies on understanding why there is an elevated level of Type I error when additive treatment effects are present. Let the parameters of the original measurements be indicated by standard Greek letters with X, Y, and k subscripts, and those of the ranked scores by Greek letters with the addition of the subscript R to denote ranked. The null hypothesis in a test of homogeneity of slopes for the original scores is  $\beta_1 = \beta_2 = \dots = \beta_k$  with each of the slopes given by

$$\beta_k = \rho_k \frac{\sigma_{Y_k}}{\sigma_{X_k}} \tag{1}$$

We dealt with the case where the null hypothesis concerning slopes implies equality of the elements on the right side of equation 1. If the null hypothesis for slopes is true then the variability of the X scores, the Y scores and the XY correlations are homogeneous. The special case where the slopes are equal because of compensating effects such as inversely related correlations and Y variances was not considered.

The question of interest concerns the equality of the  $\beta_k$ s but is tested by evaluating the null hypothesis concerning the equality of the  $\beta_{Rk}$ s. This will be an equivalent test if the terms on the right side of:

$$\beta_{Rk} = \rho_{Rk} \frac{\sigma_{RY_k}}{\sigma_{RX_k}} \tag{2}$$

are homogeneous when the terms on the right side of (1) are homogeneous.

The variances of the raw X scores ( $\sigma^2_{Xk}$ ) are homogeneous by the nature of an experiment. Under the standard procedures associated with ANCOVA the subjects are randomly assigned to conditions with no impact of treatment present in the X scores. The variances for the ranked X scores,  $\sigma^2_{RXk}$ , will be  $((kn)^2 - 1)/12$  and the sampled set of ranks from all k groups should estimate this parameter because of the random assignment. Additive treatment effects will have no impact on either  $\sigma^2_{Xk}$  or  $\sigma^2_{RXk}$ .

The correlation between the ranked XY scores ( $\rho_{Rk}$ ) will be similar but not identical to the correlation between the original scores. If the treatment conditions have equal correlations in their raw score form, that equality of correlation will be maintained in the ranked scores. Additive treatment effects should have no or only minor influences on the homogeneity of correlations based on ranked scores.

As with the ranked X scores, the variance of the ranked Y scores is a simple function of sample size (n) and number of groups (k). If there are no additive treatment effects the variance of the ranked Y scores is:

$$\sigma^2_{RY_k} = \frac{(kn)^2 - 1}{12} \tag{3}$$

Additive treatment effects have the possibility of changing the variance of ranked Y scores within a group. Since group slopes are a function of the standard deviations of X and Y, additive treatments could produce the appearance of an interaction. This possibility is most easily seen in an exaggerated example. Consider the pattern of treatment effects for 4 groups of {0, 0, 0, c} where c is an additive constant. Let c be so large that the fourth sample of scores is raised so that no member of group 4 has a score lower than the highest score in the remaining groups. In this case the ranked Y variances estimated by the first 3 of the k=4 groups would be:

$$\sigma^2_{RY_k} = \frac{((k-1)n)^2 - 1}{12} \tag{4}$$

while the variance of the fourth group would reflect the variability in ranks of  $n$  adjacent scores which is

$$\sigma^2_{RY_4} = \frac{n^2 - 1}{12} \quad (5)$$

The differences in the variability from equation 4 and 5 would produce a set of slopes in which the last group would have a slope almost  $k-1$  times smaller than the slopes of the remaining groups. For smaller additive treatment effects the separation would be less complete but still result in the reduction of the  $Y$  variability for the separated group and thus a reduction in the slope. Headrick and Sawilowsky's (2000) report of high rejection rates of the null hypothesis concerning equal slopes are Type I errors in the sense that there were only additive rather than interactive effects present. The rejections are also correct rejections of the null hypothesis concerning slopes after the additive treatment effects have confounded additive and interactive effects when  $Y$  is ranked. The proportion of rejected hypotheses will depend on the power, which is a function of the correlation and sample size.

Other configurations of additive effect would not produce the same effect. Patterns such as  $\{0, 0, c, c\}$  or  $\{-2c, -c, 0, c, 2c\}$  would alter all of the groups'  $Y$  variabilities equally and thus retain equal slopes in the ranked scores if there were equal slopes in the original distributions. In the simulations performed by Stephenson and Jacobson (1988) the vector of additive effects was  $(1, 0, 1.5, 3)$ . This pattern did not produce inflated Type I error rates as the spacing is relatively equal and the sample size and correlation were much lower than in Headrick and Sawilowsky, providing little power.

To eliminate the potential of additive treatment effects confounding the test of differences in slope we propose that the ranking of observations be based on a function of the scores that would eliminate any additive effects. The first alternative is to subtract the appropriate group sample mean from each score prior to ranking the observations and conducting the analysis of covariance. Scores within a treatment condition are defined as  $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ . The sample mean has an expected value of  $\mu + \alpha_j$ . Analysis of the

deviation from the group mean provides estimates of a common  $\varepsilon_{ij}$ .

The second alternative is to subtract the sample median prior to ranking the observations. Like the sample mean, the sample median will cancel additive treatment effects. Any constant difference reflecting the difference between the population mean and median should be eliminated when the differences are ranked. The median is offered as an alternative when the distribution of  $Y$  scores may be highly skewed.

Consider the situation in which the null hypothesis concerning slopes is true but the outcome measure is a right skewed, heavy-tailed distribution. The presence in a sample of a single, outlying score would produce deviations from the mean that were primarily negative, reflecting the inflating effect of the extreme score on the sample mean. The predominance of negative deviations along with the outlying positive deviation would distort the slope and inflate the Type I error rate.

A number of robust statistics are available to decrease the influence of extreme scores. The sample median is one of the simplest and is used in the current alternative approach. In both proposed methods the test for additive treatment effects would have to be conducted using the normal Conover and Iman (1982) or alternative method. The subtraction of either the sample mean or the sample median from scores eliminates any additive effects and precludes the deviations from being used to evaluate additive effects.

## Methodology

All simulations were conducted on a Unix computer using programs written in FORTRAN 77. Unit normal distributions were generated using the RNNOR subroutine of IMSL. All Type I error rates were obtained from 50,000 iterations of the program. For the nominal value of .05 this number of iterations produces a standard error of .001. The simulations were all based on a one-way design with  $k=4$  groups,  $n=20$  subjects, and a single covariate. Two levels of relationship between  $X$  and  $Y$  were created to represent a relatively low and relatively high degree of relationship. In the normally distributed  $X$  and  $Y$  scores the two levels represent correlations of .3 and .7.

The normally distributed covariate X was generated by RNNOR. The Y variable was created as a weighted linear combination of X and a second randomly created normal distribution to introduce random error. The weights were selected so that the variance of the Y scores was 1 and the slopes for all groups would be either .3 or .7. The original normally distributed X and Y variables (NOR X and NOR Y, respectively) were then transformed to three other shaped distributions using Fleishman's (1978) power vector method.

A platykurtic distribution was selected for study with skew of 0 and kurtosis of -1 (PLAT X and PLAT Y). The other two distribution were leptokurtic, the first with skew of 0 and kurtosis of 1.5 (LEPTO X and LEPTO Y) and finally, a more extreme, skewed, leptokurtic distribution with skew of 1.75 and kurtosis of 3.75 (SKLPT X and SKLPT Y).

All 16 possible combinations of shape of X and shape of Y were analyzed. Because of the multiple pairings no attempt was made to correct the correlations to exactly .3 and .7 in all pairings (see Headrick and Sawilowsky, 1999). The actual correlations for the 16 pairings varied from .22 to .30 for the nominal .3 and from .55 to .70 for .7. The first three shapes with no skew had much more homogeneous correlations, ranging from .28 to .30 and from .66 to .70 for .3 and .7, respectively. We shall refer to the two conditions as Low and High correlation, respectively.

Three configurations of additive treatment effects were used to evaluate the previously reported confounding of additive treatment effects with the test of slopes. The first condition had no additive effects. The second and third had configurations of 0, 0, 0, c and 0, 0, c, c, respectively. The four levels of additive constant c were .8, 1.4, 2.0, and 2.6. This produced  $1+(2)(4)=9$  distinct patterns. Because both X and Y have unit variance the additive constants are in z-scores.

Each data set was analyzed with four representations of the data. These are:

- |                       |   |
|-----------------------|---|
| 1. X- Original Scores | Y-Original Score (XY)                         |
| 2. X- Ranked Scores   | Y-Ranked Scores (RxRy)                        |
| 3. X- Ranked Scores   | Y-Ranked deviation from sample mean (RxR1y)   |
| 4. X- Ranked Scores   | Y- Ranked deviation from sample median(RxR2y) |

The analysis of the data set (XY) is the standard parametric analysis of covariance, the second (RxRy) is the Conover and Iman (1982) non-parametric analysis of covariance, the third (RxR1y) and fourth (RxR2y) are the non-parametric analyses of covariance developed in the current paper based on the mean and median, respectively.

## Results

The results were obtained by averaging the probabilities of Type I error across the simulations. The primary findings are a description of those variables that impact Type I error. In addition each Type I error rate is classified as to whether it exceeds either Bradley's (1978) conservative or liberal criterion for robustness. Although these criteria are arbitrary they provide a commonly known standard for evaluating the magnitude of the elevation of Type I error.

The first two issues deal with the relationship between the shape of the underlying distribution and Type I error. The analyses are based on the conventional analysis of covariance of the original scores, XY. Table 1 contains the mean Type I error rates for all combinations of shapes for X and Y. The results are presented separately for the low and high correlation conditions. Each mean is based on the simulations representing the nine different additive treatment combinations. Preliminary analyses indicate that additive configurations and magnitude represent trivial factors and could be combined without loss of information. Also included in Table 1 are the number of the simulations that had Type I error rates that exceeded Bradley's conservative (.055) and liberal (.075) criterion level for robustness. Only the upper limits are considered, as the present concern is for unacceptably high Type I error rates. Low error rates are more likely to be reflected in poor power.

The average Type I error rate for both LEPTO Y and SKLPT Y are considerably larger than for the normal curve. PLATY has a conservative Type I error rate. The inflated Type I error rates associated with leptokurtic curves is further seen in the frequency with which the Type I error rate exceeds even the most liberal of robustness criteria.



Table 1. Average Type I error rates across raw X and Y distributions and correlations.

Corr.	Y Dist.	X Distribution			
		PLAT X	NOR X	LEPTO X	SKLPT X
LOW $r \approx .3$	PLAT Y	.044 (0,0)	.041 (0,0)	.040 (0,0)	.043 (0,0)
	NOR Y	.050 (0,0)	.050 (0,0)	.051 (0,0)	.052 (0,0)
	LEPTO Y	.056 (7,0)	.059 (9,0)	.062 (9,0)	.062 (9,0)
	SKLPT Y	.060 (9,0)	.068 (9,0)	.072 (9,0)	.124 (9,9)
HIGH $r \approx .7$	PLAT Y	.025 (0,0)	.020 (0,0)	.034 (0,0)	.051 (0,0)
	NOR Y	.055 (6,0)	.049 (0,0)	.058 (9,0)	.068 (9,0)
	LEPTO Y	.083 (9,9)	.094 (9,9)	.103 (9,9)	.094 (9,9)
	SKLPT Y	.114 (9,9)	.178 (9,9)	.213 (9,9)	.225 (9,9)

Note. Numbers in parentheses are the number of times Type I error exceeded .055 and .075 in that condition where the maximum is 9.

The variability in the means presented in Table 1 is partitioned into the main effects and interactions between the independent variables in the simulation. Table 2 contains the mean square deviations for these sources. Because of the number of iterations all of the effects are significant based on the most conservative of standards. In the current discussion it is the relative size of the effects that is of primary concern.

Three effects are much larger than the remaining sources. These are the shape of the original Y distribution, the strength of the XY correlation, and the interaction between the shape of the Y distribution and the correlation. The shape of the X distribution has a mean square less than one-tenth that of the shape of the Y distribution. The interaction between the shapes of X and Y is small and trivial.

Table 2. Sources of variation on Type I error rate with raw X and Y scores (XY).

Source	SS	df	MS
Y Distribution shapes	0.365	3	0.122
Correlation (COR)	0.079	1	0.079
Y * COR	0.129	3	0.043
X Distribution shapes	0.034	3	0.011
X*Y	0.044	9	0.005
X * COR	0.008	3	0.003
X*Y * COR	0.012	9	0.001
Residual	0.0004	256	0.000

The main effect for the shape of Y reflects the variability in the overall means. The interaction is reflected in Type I error rates that are more extreme with a higher correlation. Platykurtic curves become more conservative and leptokurtic curves more liberal. Because there were more leptokurtic curves than platykurtic curves the average Type I error rate for the higher correlation is larger. The pattern of the means for the shapes of the X distribution mirror those of Y but are much less extreme.

The next two issues deal with the ability of ranking methods to control Type I errors for differences in slopes when there are additive treatments present. Table 3 presents the Type I error rates for the ANCOVA test of slopes proposed by Conover and Iman (1982), the ANCOVA test of slopes based on deviations of scores from the appropriate sample mean, and the ANCOVA test of slopes based on deviations of scores from the appropriate sample median. Two patterns of treatment effect, {0,0,0,c} and {0, 0, c, c} are paired with four levels of c. The results are summed across the 4x4=16 distributional pairings. Results are reported separately for low and high correlations. The parenthetical values indicate how many of these 16 simulations produced Type I error rates that exceeded the conservative and liberal robustness criteria.

Table 3. Average Type I error across correlation, treatment effect, treatment effect pattern and ranking method.

Corr.	Pattern	Data Set	Treatment Effect (c)			
			.8	1.4	2.0	2.6
LOW $r \approx .3$	0 0 0 c	RxRy	.048 (4,0)	.053 (4,0)	.061(16,0)	.068 (16,0)
	0 0 c c	RxRy	.052 (4,0)	.048 (4,0)	.047 (2,0)	.046 (0,0)
	0 0 0 c	RxR1y	.045 (4,0)	.045 (4,0)	.045 (4,0)	.045 (4,0)
	0 0 c c	RxR1y	.046 (4,0)	.045 (4,0)	.046 (4,0)	.045 (4,0)
	0 0 0 c	RxR2y	.048 (0,0)	.048 (0,0)	.048 (0,0)	.048 (0,0)
	0 0 c c	RxR2y	.048 (0,0)	.047 (0,0)	.048 (0,0)	.048 (0,0)
HIG H $r \approx .7$	0 0 0 c	RxRy	.050 (4,4)	.088(11,5)	.155(16,16)	.254 (16,16)
	0 0 c c	RxRy	.044 (4,4)	.045 (4,4)	.041 (4,3)	.034 (4,0)
	0 0 0 c	RxR1y	.031 (4,0)	.031 (4,4)	.031 (4,4)	.031 (4,4)
	0 0 c c	RxR1y	.031 (4,4)	.031 (4,4)	.031 (4,4)	.031 (4,4)
	0 0 0 c	RxR2y	.044 (4,0)	.044 (4,0)	.043 (4,0)	.043 (4,0)
	0 0 c c	RxR2y	.044 (4,0)	.044 (4,0)	.043 (4,0)	.043 (4,0)

Note. Numbers in parentheses are the number of times Type I error exceeded .055 and .075 in that condition where the maximum is 16. Rx indicates ranked X scores. Ry indicates ranked Y scores. R1y indicates ranked deviations of Y from the group Y mean. R2y indicates ranked deviations of Y from the group Y median.

The Type I error rate for RxRy increases as the magnitude of the treatment effect increases for the {0,0,0,c} pattern but not for the {0,0,c,c} pattern. The corresponding values for the methods based on deviations, RxR1y and RxR2y, have mean Type I error rates near .05. The simulations based on deviation scores with Type I error rates that surpassed the conservative robustness criteria are those based on SKLPT Y. The effects are more pronounced when there is a high correlation than when there is a low one.

The two new methods perform similarly in most of the simulations. The difference between them is predicted to be when there is a very skewed distribution. Table 4 presents the average Type I error rate of the {0,0,0,c} pattern for LEPTO Y and SKLPT Y distributions. The results are summed across shape of the X distribution and the additive constants. As expected the Type I error rate for the method based on deviations from the mean became problematic when the distribution is skewed. A symmetric leptokurtic distribution showed no elevation of Type I error with either of the new methods. The method based on the median is generally within acceptable bounds although it has more than a .06 error rate with the Skewed Leptokurtic, SKLPT Y.

Table 4. Comparing the two ranking alternatives across correlation, treatment effect and Y distribution.

	Correlation			
	LOW $r \approx .3$		HIGH $r \approx .7$	
	LEPTO Y	SKLPT Y	LEPTO Y	SKLPT Y
RxR1y	.044 (0,0)	.059 (32,0)	.023 (0,0)	.105 (36,36)
RxR2y	.044 (0,0)	.052 (0,0)	.021 (0,0)	.063 (36, 0)

Note. Numbers in parentheses are the number of times Type I error exceeded .055 and .075 respectively where the maximum is 36. Rx indicates Ranked X scores. R1y indicates Ranked Y deviations from the group Y mean. R2y indicates Ranked Y deviations from the group Y median.

### Conclusion

Both of the problems associated with conducting a test of differences in slopes were replicated in the present study. Analysis of covariance on scores

that are not normally distributed have Type I error rates that systematically vary from the nominal value. If the distribution is leptokurtic the Type I error rate will be liberal and if it is platykurtic it will be conservative. It is difficult to determine if skew plays a role as most skewed distributions are also leptokurtic. The effect of shape is most clearly present when there is considerable shared variation in X and Y.

It is clearly the shape of the outcome measure rather than the covariate that results in manipulation of the Type I error rate. There is a small effect for the shape of X and little interaction between the shapes of X and Y. The complete set of 16 shapes is probably unrealistic in real world settings. The shapes of both X and Y are likely to be related to underlying characteristics of the sample chosen so that if Y is leptokurtic then X will likely also be somewhat leptokurtic. This would result in an accumulation of the major impact of the leptokurtic Y-scores and the minor impact of leptokurtic X scores to produce even more extreme Type I error elevation.

Tests of significance involving ranks rather than the original scores largely control the inflated Type I error rate although there appear to be unexplained differences in the error rate associated with ranking methods as a function of the underlying distribution. Specifically, the error rate is consistently higher for the Conover and Iman (1982) method when the SKLPT Y distribution was the source of the ranks. This trend for skewed distributions to produce larger Type I error rates even after being ranked was also found in Conover and Iman (1982) and Stephenson and Jacobson (1988).

The influence of additive treatment effects is shown to have the potentially serious inflation of Type I error noted by Headrick and Sawilowsky (2000). The effect was found where the additive effects tended to isolate one treatment group away from the remaining groups. Since the variance of ranks is based on the range of the ranks within the complete set, the separation of one group from a set of other groups will reduce the range and variance and produce a reduced slope. The effect appeared as the magnitude of the additive treatment effect increased. The beginning additive constant of .8 corresponds to a large effect in Cohen's (1988) terms. This effect showed no inflation of Type I error rate. Only as the additive

effect increased beyond this did the error rate become problematic.

Both of the proposed methods for testing slopes in the presence of potential additive effects reduced the Type I error rate to a generally acceptable level. The simulations that resulted in somewhat higher error rates were those with the most extreme distribution SKLPT Y. The method using deviations from the sample median was superior in controlling Type I error with SKLPT Y but was still somewhat elevated.

The two tests developed differ from others in that they are solely for testing the differences in slopes. There is no companion test for the presence of additive effects. A separate test such as that in Conover and Iman (1982) would need to be used for additive effects.

The ranking methods developed herein will have to be compared to other options to determine whether they have sufficient power to replace the traditional methods. The level of additive treatment effect used in the simulation is large and, at the upper end, may represent a level seen in relatively few experiments.

The experimenter should be able to anticipate this magnitude of effect. If the analysis of simple ranked scores as proposed by Conover and Iman (1982) is more powerful than the methods based on deviations the experimenter may choose to use simple ranks unless there is the expectation that very large additive treatment effects exist. However, if the power is equivalent the methods proposed herein should be preferred as they have more general Type I error control.

Lastly, the power of the tests using deviations from the mean and median need to be compared. While the median based method has superior Type I error control with the skewed leptokurtic distribution if it has less power the researcher may again want to determine if that condition within the outcome measure is likely to be present in the data and select accordingly.

#### References

- Atiquallah, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika*, 51, 365-372.
- Bradley, J. C. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.

- Burnett, T. D., & Barr, D. R. (1977). A nonparametric analogy of analysis of covariance. *Educational and Psychological Measurement*, 37, 341-348.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2<sup>nd</sup> ed.). Hillsdale, NJ: Erlbaum.
- Conover, W. J., & Iman, R. L. (1982). Analysis of covariance using the rank transformation. *Biometrics*, 38, 715-724.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671-684.
- Cronbach, L. J. & Snow, R. E. (1981). *Aptitudes and instructional methods: A handbook for research on interactions*. (2<sup>nd</sup> ed.). New York: Irvington.
- Dance, K. A., & Neufeld, R. W. J. (1988). Aptitude-Treatment Interaction research in the clinical setting: A review of attempts to dispel the "patient uniformity" myth. *Psychological Bulletin*, 104(2), 192-213.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521-532.
- Harwell, M. R., & Serlin, R. C. (1988). An experimental study of a proposed test of non-parametric analysis of covariance. *Psychological Bulletin*, 104, 268-281.
- Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, 64, 25-35.
- Headrick, T. C., & Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics- Simulation and Computation*, 29, 1059-1087.
- Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. NY: Wiley.
- Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86(4), 721-735.
- Klockars, A. J., & Beretvas, S. N. (2001). Analysis of covariance and randomized block design with heterogeneous slopes. *The Journal of Experimental Education*, 69, 393-410.
- Klockars, A. J., & Moses, T. (2001). *Dealing with violations of the assumptions for aptitude X treatment interactions in ANCOVA*. Paper presented at the American Educational Research Association, Seattle.
- Klockars, A. J., Potter, N. S., & Beretvas, S. N. (1999). Power to detect additive treatment effects with randomized block and analysis of covariance designs. *The Journal of Experimental Education*, 67, 180-191.
- McClelland, G. H. & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376-390.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Puri, M. L., & Sen, P. K. (1968). Analysis of covariance based on general rank scores. *Annals of Mathematical Statistics*, 40, 610-618.
- Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association*, 62, 1187-1200.
- Rheinheimer, D. C., & Penfield, D. A. (2001). The effects of type I error rate and power of the ANCOVA F test and selected alternatives under nonnormality and variance heterogeneity. *The Journal of Experimental Education*, 69, 373-391.
- Seaman, S., Algina, J., & Olejnik, S. F. (1985). Type I error probabilities and power of the rank and parametric ANCOVA procedures. *Journal of Educational Statistics*, 10, 345-367.
- Shirley, E. A. C. (1981). A distribution-free method for analysis of covariance based on ranked data. *Applied Statistics*, 30, 158-162.
- Stephenson, W. R., & Jacobson, D. (1988). A comparison of non-parametric analysis of covariance techniques. *Communication in Statistics: Simulation and Computation*, 26, 605-618.

## Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When $\sigma_1^2 \neq \sigma_2^2$

Shlomo S. Sawilowsky  
Educational Evaluation and Research  
Wayne State University

---

The history of the Behrens-Fisher problem and some approximate solutions are reviewed. In outlining relevant statistical hypotheses on the probable difference between two means, the importance of the Behrens-Fisher problem from a theoretical perspective is acknowledged, but it is concluded that this problem is irrelevant for applied research in psychology, education, and related disciplines. The focus is better placed on “shift in location” and, more importantly, “shift in location and change in scale” treatment alternatives.

Key words: Behrens-Fisher problem, t test, heterogeneous variances.

---

### Introduction

Simply stated, the Behrens-Fisher problem arises in testing the difference between two means with a t test when the ratio of variances of the two populations from which the data were sampled is not equal to one. This condition is known as heteroscedasticity, which is a violation of one of the underlying assumptions of the t test. The resulting statistic is not distributed as t, and therefore the associated p values based on the entries found in standard t tables are incorrect. Use of tabulated critical values may lead to increased false positives, which are known as Type I errors, or a conservative test that lacks statistical power to detect significant treatment effects.

### Development of Student's Distribution For a Unique Sample

Regarding the development of the t test, Fisher (1939) noted,

---

Shlomo S. Sawilowsky is Professor of Educational Evaluation & Research, & Wayne State University Distinguished Faculty Fellow. His current areas of interest are nonparametric and computer intensive methods, the revival of classical measurement theory, and a recommitment to experimental design in lieu of quasi-experimental design. Email: shlomo@wayne.edu.

To the present generation of statisticians, familiar with ‘Student’s’ distribution..., it has for some time appeared to be a somewhat puzzling historical fact that this advance in simple statistical procedure was not made long before, and was not made rather by a mathematician than a research chemist.

Light is perhaps thrown on this puzzle by the contrast, which has been striking during the last twenty years, between the facility, confidence, and skill with which the new tests have been applied by practical men in research departments, and the embarrassment and confusion of many discussions, in journals devoted to mathematical statistics, by mathematically minded authors lacking contact with practical research (p. 141).

Prior to ‘Student’ or W. S. Gosset, the mathematician Helmert was able to determine the distribution of the sum of squares  $\sum(x - \mu)^2$  (Helmert, 1875) and  $\sum(x - \bar{x})^2$  (Helmert, 1876), but indicated no practical value for the results. Subsequent to Gosset, another mathematician, Burnside (1923), used Bayesian methods in rediscovering the t distribution, although the

inclusion of an a priori distribution for a precision constant resulted in a difference of one degree of freedom. Interestingly, he presented a table of quartiles of the t distribution, prompting Fisher (1941) to remark, “It evidently did not occur to him that a 5 or 1% table would be more useful...[this] may be taken to indicate that he regarded his solution rather as a matter of academic interest than as meeting a need for guidance in practical decisions” (p. 142).

According to Jeffreys (1937), the t distribution was not discovered earlier because it “involves an unstated assumption” (p. 48) that for the sample mean ( $\bar{x}$ ), estimated variance of the mean ( $s^2$ ), and population mean ( $\mu$ ), then the distribution of

$$t = \frac{\bar{x} - \mu}{s} \tag{1}$$

depends only on the sample size n. Fisher (1941) added that novel reasoning also left unstated by Gosset was that  $\bar{x}$  and  $s^2$  should be unbiased.

The question of bias in  $s^2$  was troublesome indeed. The prepublication title of “The Probable Error of a Mean” (Student, 1908) was “On the Probable Error of a Unique Sample”. The uniqueness that worried Gosset was the requirement that  $s^2$  be unbiased. Although Gosset’s paper pertained to the difference distribution of paired observations, Fisher (1941) extended this concern to the two independent samples case. Fisher suggested that one of the “difficulties in the way of an early discovery of ‘Student’s’ test” was because of “the application of the same methods to the more intricate problem of the comparison of the means of samples having unequal variances, or more correctly from populations, of which the variance ratio is unknown, and itself constitutes one of the parameters which require to be ‘Studentized’”(1941, p. 146).

The Behrens-Fisher Problem

The first expression and solution to this problem was by Behrens (1929), and reframed by Fisher (1939a) from a Fisherian perspective as

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{2n_1 + 1} + \frac{s_2^2}{2n_2 + 1}}} \tag{2},$$

where  $s_1$  and  $s_2$  are fixed and  $\sigma_1$  and  $\sigma_2$  have fiducial distributions. Tables of critical values were given in Fisher and Yates (1957). This solution was challenged by Bartlett (1936) on the principle of inverse probability from a Bayesian perspective. Fisher responded with his usual tenacious and acrid style: “From a purely historical standpoint it is worth noting that the ideas and nomenclature for which I am responsible were developed only after I had inured myself to the absolute rejection of the postulate of Inverse Probability” (1937a, p. 151; see also 1937b, 1939b). Jeffreys (1940) restored calm by demonstrating that Bartlett’s perspective was not a challenge to the Fisherian approach, but rather was another way of starting with the same hypothesis and ending with the same conclusion.

Commonly available solutions implemented in computer software statistics packages have eschewed both of those approaches in favor of a third theoretical perspective. This is the frequentist approach of Neyman-Pearson, where  $\sigma_1$  and  $\sigma_2$  are fixed, but  $s_1$  and  $s_2$  are free to vary in (2). The typical solution in statistics packages for solving the two sample problem ( $k = 2$ ) is the Welch separate variances test, which has become known as the Welch-Aspin test with modified degrees of freedom, given by

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{3}.$$

(Welch, 1937, 1949a, 1949b; Satterthwaite, 1941, 1946; Aspin 1948, 1949). Although the exact distribution of the Welch statistic is known under normality (Ray & Pitman, 1961), it remains an approximate solution to the Behrens-Fisher problem. Welch (1947) also provided a solution for the generalized problem ( $k \geq 2$ ).

The Behrens-Fisher problem continued to attract the attention mathematical statisticians and applied researchers. For example, different perspectives were given by Wald (1955), Banerjee (1960), and Pagurova, (1968). These are but a few of the many solutions published in the literature.

### Robustness With Respect To Unequal $n$ 's and Population Normality

Eventually, however, questions arose on the robustness with respect to Type I errors for unequal  $n$ 's. Fisher (1939a) tried to quash this line of research by restating the fact that Gosset's paper (Student, 1908) was on pairs of measurements (height vs length of middle finger for 3,000 criminals), obviating the unequal  $n$  problem. Nevertheless, in the context of  $k \geq 2$  independent samples, studies indicated that the various solutions were not robust to unequal  $n$ 's (e.g., Kohr, 1970; Mehta & Srinivasa, 1970; Kohr & Games, 1974; Tomarkin & Serlin, 1986). Solutions to the unequal  $n$  situation appeared which preserved nominal alpha (e.g., Scheffé, 1943; McCullough, Gurland, & Rosenberg, 1960), although some of them were subsequently found to be not very powerful.

This line of research was soon overshadowed by the concern of robustness with respect to Type I errors for departures from population normality. Monte Carlo studies showed that the Behrens-Fisher, Bartlett, and Welch-Aspin/Satterthwaite approximate solutions are not robust to departures from normality (e.g., James, 1959; Yuen, 1974). A similar fate awaited many of the other solutions, such as the Brown & Forsythe (1974) test (Clinch & Keselman, 1982), and the  $H_m$  test by Wilcox (1990) which had "the tendency to be conservative" (Oshima & Algina, 1992, p. 262) for long-tailed distributions. The inability of these procedures to maintain the Type I error rate at nominal alpha created the opportunity for another round of alternative solutions being published.

Some solutions based on nonparametric or nonparametric-like procedures were unsuccessful. For example, Pratt (1964) showed that the Mann-Whitney U (Mann & Whitney, 1947) and the expected normal scores test (Hájek & Sidák, 1967) resulted in nonrobust Type I error rates. Bradstreet (1997) found the rank transform test (Conover & Iman, 1982) to result in severely inflated Type I error rates. For the case of  $k > 2$ , Feir-Walsh and Toothaker (1974) and Keselman, Rogan, and Feir-Walsh (1977) found the Kruskal-Wallis test (Kruskal & Wallis, 1952) and expected normal scores test (McSweeney & Penfield, 1969) to be "substantially affected by inhomogeneity of variance" (p. 220).

Other nonparametric solutions met with more success. Yuen (1974) provided a robust solution based on trimmed means and matching sample variances. Tiku and Singh's (1981) solution was based on modified maximum likelihood estimators. Tan and Tabatabai (1985) combined the Tiku and Singh procedure with the Brown-Forsythe test to produce a more powerful procedure than those based only on Huber's M estimator (Huber, 1981; Schrader & Hettmansperger, 1980).

The development of procedures involving the Behrens-Fisher problem is not restricted to the usual  $k \geq 2$  independent samples case. Games and Howel (1976) examined pairwise multiple comparison solutions. Bozdogan and Rameriz (1986) proposed a likelihood ratio for situations where only subsets respond to a treatment. Johnson and Weerahandi (1988) provided a Bayesian solution to the multivariate problem. Koschat and Weerahandi (1992) developed a class of tests for the problem of inference for structural parameters common to several regressions.

Despite the many approximate solutions published to date, the Behrens-Fisher problem remains actively studied. In the past 35 years, there were 37 doctoral dissertations completed pertaining to some aspect of the Behrens-Fisher problem, including newly proposed approximate solutions (*Dissertation Abstracts Online*, 2000). There was one dissertation completed in the 1960s, six in the 1970s, 16 in the 1980s, and 14 in the 1990s.

### Hypothesis Testing

Consider the entries in Table 1. It contains the various hypotheses on the probable error of a mean, and the probable difference between two means. Hypotheses #1-#3 rarely occur in applied studies because they pertain to the Z test which requires  $\sigma^2$  to be known. It is unusual for a social and behavioral science researcher to have the entire population at her or his disposal, or to know the parameters of the population. Z tests are valuable mainly as a pedagogical tool for introducing inferential statistics to students of data analysis methods.

Table 1. Parametric Nondirectional (Two-Sided) Null ( $H_0$ ;) And Alternative ( $H_a$ ;) Hypotheses For One Sample ( $\mu_0$ ) And Two Samples ( $\mu_1, \mu_2$ ) Z And t Tests.

---

Z tests: Hypotheses That Rarely Occur In Applied Studies

- #1:  $H_0: \mu_1 = \mu_0; \sigma^2$  is known  
 $H_a: \mu_1 \neq \mu_0; \sigma^2$  does not change
- #2:  $H_0: \mu_1 = \mu_2; \sigma_1^2 = \sigma_2^2$  and known  
 $H_a: \mu_1 \neq \mu_2; \sigma_1^2$  and  $\sigma_2^2$  do not change
- #3:  $H_0: \mu_1 = \mu_2; \sigma_1^2 \neq \sigma_2^2$ , but known  
 $H_a: \mu_1 \neq \mu_2; \sigma_1^2$  and  $\sigma_2^2$  do not change

t tests: Hypotheses That Occur In Applied Studies - The "Shift in Location Alternative"

- #4:  $H_0: \mu_1 = \mu_0; \sigma^2$  is unknown, but assumed to be unbiased  
 $H_a: \mu_1 \neq \mu_0; \sigma^2$  does not change
- #5:  $H_0: \mu_1 = \mu_2; \sigma_1^2$  and  $\sigma_2^2$  are unknown, but assumed to be equal  
 $H_a: \mu_1 \neq \mu_2; \sigma_1^2$  and  $\sigma_2^2$  do not change

The Two Sample Behrens-Fisher Problem (Fisherian & Bayesian)

- #6a:  $H_0: \mu_1 = \mu_2; \sigma_1^2$  and  $\sigma_2^2$  are unknown, but it is known that  $\sigma_1^2 \neq \sigma_2^2$
- #6b:  $H_0: \mu_1 = \mu_2; \sigma_1^2$  and  $\sigma_2^2$  are unknown, but cannot be assumed to be equal

The Two Sample Behrens-Fisher Problem (Neyman-Pearson)

- #6c:  $H_0: \mu_1 = \mu_2; \sigma_1^2$  and  $\sigma_2^2$  are unknown, but it is known that  $\sigma_1^2 \neq \sigma_2^2$   
 $H_a: \mu_1 \neq \mu_2; \sigma_1^2$  and  $\sigma_2^2$  do not change
- #6d:  $H_0: \mu_1 = \mu_2; \sigma_1^2$  and  $\sigma_2^2$  are unknown, but cannot be assumed to be equal  
 $H_a: \mu_1 \neq \mu_2; \sigma_1^2$  and  $\sigma_2^2$  do not change

Hypotheses That Frequently Occur in Applied Studies: The "Shift in Location and Change in Scale" Alternative

- #7:  $H_0: \mu_1 = \mu_2$  and  $\sigma_1^2 = \sigma_2^2$   
 $H_a: \mu_1 \neq \mu_2$  and  $\sigma_1^2 \neq \sigma_2^2$

---

*Note:*  $H_a$ : can be expressed as a directional (one-sided) hypothesis by replacing " $\neq$ " with either " $>$ " or " $<$ ".

Hypotheses #4 and #5 refer to the "shift in location" alternative and are tested by the t test. Although no test can survive violations of independence of observations, under certain commonly occurring conditions (i.e., sample sizes are equal or nearly so and are at least 25 to 30, and tests are two-tailed rather than one-tailed), the t test is remarkably robust with respect to both Type I and II errors for departures from normality (e.g., Sawilowsky, 1990; Sawilowsky & Blair, 1992).

Editors and reviewers challenge the shift alternative as a realistic treatment outcome, which in turn, questions the applicability of Hypotheses #4 and #5 to real world data sets. After studying the histograms of many real treatment vs control and pretest-posttest data sets, I argue that, indeed, shift happens. An example with 714 admit vs discharge Functional Independence Measure scores (Keith, Granger, Hamilton, & Sherwin, 1987), an instrument that is frequently used in the field of rehabilitation counseling, was shown in Nanna and Sawilowsky (1998).

(I would be remiss if I failed to note that numerous Monte Carlo studies have shown that the nonparametric Wilcoxon Rank Sum test can be three to four times more powerful in detecting differences in location parameters when the normality assumption was violated (e.g., Blair & Higgins, 1980a, 1980b, 1985; Blair, Higgins, & Smitley, 1980; Sawilowsky & Blair, 1992). Micceri (1989) found that only about 3% of real data sets in psychology and education are relatively symmetric with light tails. Therefore, the Wilcoxon procedure should be the test of choice. The t test remains a popular test, however, most likely due to the inertia of many generations of classically parametrically trained researchers who continue its use for this situation.)

As noted by #6a - #6d, the hypotheses tested by the Behrens-Fisher problem can be expressed from the Fisherian/Bayesian perspective by the absence of an alternative hypothesis, or in the Neyman-Person frequentist paradigm. In the first example according to both perspectives (i.e., #6a and #6c), it is known that samples were drawn from two different populations (e.g., the first may have been extreme asymmetric such as exponential decay and the second may have been multimodal from a likert scale), but the population parameters remain unknown. Thus, the Behrens-Fisher problem arises because the ratio of



population variances is different from one, although neither constituent value is known. The second and more common example, according to both perspectives (i.e., #6b and #6d), indicates that no information is available on the population from which the samples were drawn, and it cannot be safely assumed that the ratio of population variances is equal to one. Now, I discuss two reasons why these situations are important, and two reasons why they are irrelevant to applied researchers.

### Two Reasons Why The Behrens-Fisher Problem Is Important

1. The Behrens-Fisher problem is a classic. Many prestigious mathematical statisticians and applied researchers have addressed this problem. For some, their careers began with this problem; for others, their careers ended with this problem. The Behrens-Fisher problem has as much mystique and has received as much fanfare in its discipline as other classical problems that remain unsolved or unfinished in their disciplines, such as these:

- In 1630, Pierre de Fermat, an amateur mathematician, wrote “hanc marginis exiguitas non caperet” - he found a proof that was too large to write in a marginal note in his copy of the ancient Greek Diophantus’ *Arithmetica* that  $x^n + y^n = z^n$  has no nonzero integer solutions for  $x$ ,  $y$  and  $z$  when  $n > 2$ . In October, 1994, the mathematician Andrew Wiles solved the final aspect of this conjecture. (Fermat’s last conjecture is a special case of  $x^n + y^n = cz^n$ , which remains unproven.) However, Wiles noted, “Fermat couldn’t possibly have had this proof. It’s a 20th-century proof. There’s no way this could have been done before the 20<sup>th</sup>-century” (Wiles, 1996). Thus, the conjecture remains unproven using 17<sup>th</sup> century mathematics.
- In 1822, Franz Schubert wrote what was later to be known as the ‘Unfinished’ Symphony No. 8 (or No. 7 according to some numbering schemes) in B Minor. He worked on it for six years, but only completed the first two movements of an

intended four movement symphony. Mysteriously and uncharacteristically, he moved on to other pieces without finishing this symphony. Many musicians have written what they imagine the final two movements might have been if Schubert had finished it.

- In the 20<sup>th</sup> Century, physicists theorized on the unification of the laws of the universe. However, the solution eluded physicists from Albert Einstein to Stephen Hawking. (The so-called “Grand Unification Theories” combine the weak, strong, and electromagnetic forces, but leave out gravity.)

2. The second reason that the Behrens-Fisher problem is important is due to the byproducts that have been developed in the course of creating approximate solutions. Some examples include:

- Bartlett’s (1937) study of heteroscedasticity culminated in a well known Chi-Squared test on variances, which is useful for testing the underlying assumption of homoscedasticity. Bartlett’s test is a logarithmic modification of the Neyman and Pearson (1931)  $L_1$  test for the equality of variances of  $k$  groups.
- James’ (1959) attempt to improve on the Behrens-Fisher, Welch, and Yates (1939) solutions led to the development of a Cornish-Fisher expansion for a symmetric distribution.
- Statistics were developed throughout the 20<sup>th</sup> Century based on asymptotic or large sample theory. Many were published based on elegant mathematical statistical theory, but turned out to be invalid for use in applied work. The Behrens-Fisher problem highlighted the importance of conducting robustness and comparative power studies relative to small samples.

(Regarding the last point, my recommendation is that authors of new statistics or procedures publish their work *after* they have

conducted studies on the properties of the statistic when underlying assumptions are violated. Note that further study is moot if results for expedient mathematical distributions produce poor results; but if good results are obtained, verification is still required with real data sets.)

Two Reasons Why The Behrens-Fisher Problem Is Irrelevant

1. Howell and Games (1974) suggested that “Educational and psychological researchers often deal with groups that tend to be heterogeneous in variability” (p. 72). This is mitigated by the fact that, “We have spent many years examining large data sets but have never encountered a treatment or other naturally occurring condition that produces heterogeneous variances while leaving population means exactly equal” (Sawilowsky and Blair, 1992, p. 358). None of Micceri’s (1989) 440 real psychology and education data sets reflected this condition, nor have I seen an example in the literature. Thus, the issue of heterogeneous variances and their impact on Type I errors is moot.

Zumbo and Coulombe (1997) demurred, and claimed “We could simply counter that in our experience we have seen it occur” (p. 148), but there was no data set in their article. Algina and Olejnik (1984) referred to a data set in Box and Cox from 1964, but the reference is missing from their bibliography. The ratios of minimum (0.0001) to maximum (0.1131) variances given for the 12 entries in their 3x4 layout are impressive; the frequency with which psychological and educational instruments produce variances less than one-twelfth of a single point remains problematic. Koschat and Weerahandi (1992) refer to what appears to be a real data set from business and economics, although they only published summary statistics and not the actual data set. Even if examples can be found, the question remains if the Behrens-Fisher problem surfaces with such frequency that merits the journal space it has been given.

2. The most prolific treatment outcome in applied studies is known. It is where a change in scale is concomitant with a shift in means. As an intervention is implemented, the means increase or decrease according to the context. Simultaneously, the treatment group may become more homogeneous on the outcome variable due to

sharing the same intervention, method, conditions, etc. Alternatively, the group may become more heterogeneous, as some respond to the treatment while others do not respond, or even regress.

What Is Wrong With Testing For Homogeneity Prior To The t-Test?

A common strategy is to conduct a test on variances prior to the pooled samples t test (e.g., SAS, 1990, p. 25; SPSS, 1993, p. 254-255; SYSTAT, 1990, p. 487). If the F test on variances, for example, is not significant, then the researcher continues with the t test. However, if the F test is significant, then the researcher is advised to conduct the separate variances t test (e.g., Welch-Aspin) with modified degrees of freedom.

There is a serious problem with this approach that is universally overlooked. The sequential nature of testing for homogeneity of variance as a condition of conducting the independent samples t test leads to an inflation of experiment-wise Type I errors. A small Fortran program was written, compiled, and executed to demonstrate this, with the results noted in Table 2.

Table 2. Type I Error And Power For The Pooled-Variates Independent Sample t-test Conducted Unconditionally Or Conditionally On The F Test For Homogeneity Of Variance,  $\alpha = 0.050$ ;  $n_1 = n_2 = 5$ , 100,000 Repetitions.

Distribution	t-test				F-test Type I Error
	Unconditional		Conditional		
	L	R	L	R	
Normal					
c=0.0	.025	.025	.023	.023	.051
c=0.95	.000	.265	.000	.252	
c=2.0	.000	.790	.000	.750	
Chi-Square (v=2)					
c=0.0	.023	.019	.015	.013	.172
c=1.5	.000	.252	.000	.202	
c=3.5	.000	.735	.000	.632	

Note: “c” = shift in location to produce approximately small or large Effect Sizes. A study of robustness with respect to Type II errors requires “c” to represent equal Effect Sizes across distributions, which was not done for this illustration. “L” = left tail. “R” = right tail.

An examination of Table 2 highlights a number of important points:

- The experiment-wise Type I error rate, under normality, is .097 (.051+.023+.023) when the t test is conducted conditional on the F test for homogeneity of variance. This is almost twice nominal alpha.
- The experiment-wise Type I error rate when the data were sampled from a Chi-Squared distribution ( $v=2$ ) is .200, which is four times nominal alpha!
- The F test on variances, as is well known, is nonrobust to departures from normality. In this case the Type I error rate for Gaussian data of 0.051 ballooned up to .172 for the Chi-Squared ( $v=2$ ) data. This inflation level of about 3.5 times nominal alpha means the data analyst will frequently abandon the pooled samples t test in favor of the separate variances test, when in fact, the condition of homoscedasticity holds. This problem can be ameliorated somewhat by using Levene's (1960) test, which is more robust to departures from normality.
- Conducting the t-test conditioned on the F test for variances resulted in a 5% loss of power under normality, which is ill afforded in small samples applied research.
- Conducting the t-test conditioned on the F test for variances resulted in a 20% loss of power under the Chi-Squared ( $v=2$ ) distribution for the small Effect Size, and a 14% loss in power for the large Effect Size, which is ill afforded in small samples applied research.

Hyman (1995) opined that methodology articles are less helpful when they are restricted to pointing out errors or deficiencies, and are more helpful when they redirect researchers toward a useful methodology. Given the severity of the problem of pursuing Hypothesis #6 sequentially after a test on variances, it is appropriate to review Hypothesis #7 in more detail.

#### Refocusing On Treatments That Impact Location And Scale

Hypothesis #7 pertains to the situation where naturally occurring differences or treatment outcomes produce a shift in location and a change in scale. Diamond (1981, p. 73-74) discussed a simple procedure where variances and means are tested separately. What is needed, however, is a test of both parameters simultaneously. Lepage (1971, 1975), Gastwirth and Podgor (1992), and Podgor and Gastwirth (1994) offered some early work and hypothesis tests that depend on location and scale. Two more recently developed statistics for Hypothesis #7 were given by O'Brien (1988) and Brownie, Boos, and Hughes-Oliver (1990). They are discussed below because they are promising for small samples applied research.

(1) O'Brien's (1988) generalized t-test is carried out by ordinary least squares or logistic regression. In terms of the former, a dummy variable of 1, representing group membership, or 0, representing nonmembership, is regressed on the outcome variable,  $w$ , as well as  $w^2$ :

$$y' = \beta_0 + \beta_1 w + \beta_2 w^2 \quad (4).$$

If  $\beta_2$  is not near zero, the test for treatment effects is conducted with the 2 degrees of freedom F test of  $H_0: \beta_1 = \beta_2 = 0$ . If  $\beta_2$  is near 0, however, (4) is replaced with

$$y' = \beta_0 + \beta_1 w \quad (5),$$

and the one degree of freedom test of  $H_0: \beta_0 = 0$ , an independent samples t test, is conducted. It is called a generalized t-test because of the variety of levels of nominal  $\alpha$  which may be selected for testing (4).

Blair and Morel (1991) examined the experiment-wise Type I error rate of conducting (5) conditional on (4). The sequential conditional testing procedure resulted in inflated Type I errors. Grambsch and O'Brien (1991) provided a "2/3" rule, where approximately correct Type I errors are obtained by reducing alpha to two-thirds of the desired size. Subsequently, a superior solution was made available by Blair (1991), who provided a corrected table of critical values for O'Brien's procedure which results in correct Type I error rates.

(2) Brownie, Boos, and Hughes-Oliver (1990) provided a modification to the t test:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{s_1^2 \sqrt{\frac{1}{n_1} \times \frac{1}{n_2}}} \quad (6),$$

where  $s_1^2$  is the sample variance from the control group, and  $v = n_1 - 1$ . Subsequently, Sawilowsky et al. (1991) and Blair and Sawilowsky (1993a, 1993b) demonstrated through Monte Carlo methods that  $t^*$  is not robust with respect to Type I errors for departures from population normality. In addition, it requires that the change in scale increase, but not decrease. Blair and Sawilowsky (1992a, 1992b, 1993a, 1993b) fixed the Type I error properties by developing two new tests based on  $t^*$  and  $F^*$ , the extension based on  $k > 2$ . In the context of  $F^*$ , the first test is a permutation analogue ( $pF^*$ ), which does not require a priori knowledge of the expected change (i.e., increase or decrease) in variability relative to the control groups.

The second ( $pF^*_{\min}$ ) designates the group with the smallest variance as the control group, and substitutes  $s_{\min}^2$  for  $s_1^2$  in (6). (Both procedures can also be conducted as an approximate randomization test with negligible loss in precision or power.) These tests and other procedures were examined further by Troendle, Blair, Rumsey, and Moke (1997).

Podgor and Gastwirth (1994) compared O'Brien's test with Brownie, Boos, Hughes-Oliver's test in various configurations. However, they did not use Blair's corrected critical values or Blair and Sawilowsky's approximate randomization correction. One of my doctoral students is comparing both procedures with their respective corrections with two nonparametric tests. One statistic is the Savage test for positive random variables (which received some attention by Podgor & Gastwirth, 1994). It assumes that a difference in scale causes a difference in location (see, e.g., Deshpande, Gore, & Shanubhogue, 1995, p. 53-56). The other is the Rosenbaum test for general differences (see, e.g., Neave & Worthington, 1988, p. 144-149).

## Conclusion

The Behrens-Fisher problem is a classic, but its many and continuing solutions are perhaps better housed in journals catering to theoretical developments. Sufficient journal space has been given to this problem in comparison with the frequency with which it occurs. Instead, applied researchers should focus on more practical treatment outcomes, such as a treatment or naturally occurring condition that brings about a shift in location and a change in scale. This is the most realistic treatment outcome in applied psychology and education research. It presents an exciting area in which considerable additional research is warranted.

## References

- Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and Psychological Measurement, 44*, 39-48.
- Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika, 35*, 88-96.
- Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika, 36*, 290-296.
- Banerjee, S. K. (1960). Approximate confidence intervals for linear functions of means of  $k$  populations when the population variances are not equal. *Sankhyā, 22*, 357-358.
- Bartlett, M. S. (1936). The information available in small samples. *Proceedings of the Cambridge Philosophical Society, 32*, 560-566.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Royal Society of London Proceedings, Series A, 160*, 268-282.
- Behrens, W. -V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher, 68*, 807-837.
- Blair, R. C. (1991). New critical values for the generalized t and generalized rank-sum procedures. *Communications in Statistics, 20*, 981-994.

- Blair, R. C., & Higgins, J. J. (1980a.) A comparison of the power of the t test and the Wilcoxon statistics when samples are drawn from certain mixed normal distributions. *Evaluation Review*, 4, 645-656.
- Blair, R. C., & Higgins, J. J. (1980b). A comparison of the power of the Wilcoxon's rank-sum statistic to that of student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, 5, 309-335.
- Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, 97, 119-128.
- Blair, R. C., & Morel, J. G. (1991). On the use of the generalized t and generalized rank-sum statistics in medical research. *Statistics in Medicine*, 11, 491-501.
- Blair, R. C., & Sawilowsky, S. S. (1992a). A comparison of the generalized and modified t tests. Annual meeting of the American Educational Research Association, SIG Educational Statisticians, San Francisco, CA.
- Blair, R. C., & Sawilowsky, S. S. (1992b). Type I error and power of the modified and generalized t tests. 1992 Abstracts: Joint Statistical Meetings of the American Statistical Association, Biometrics Society, and Institute of Mathematical Statistics. Boston, MA, p. 49.
- Blair, R. C., & Sawilowsky, S. S. (1993a). A note on the operating characteristics of the modified F test. *Biometrics*, 49, 935-939.
- Blair, R. C., & Sawilowsky, S. S. (1993b). Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls. *Statistics in Medicine*, 12, 2233-2243.
- Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.
- Bozdogan, H., & Ramirez, D. E. (1986). An adjusted likelihood-ratio approach to the Behrens-Fisher test. *Communications in Statistics*, 15, 2405-2433.
- Bradstreet, T. E. (1997). A Monte Carlo study of type I error rates for the two-sample Behrens-Fisher problem with and without rank transformation. *Computational Statistics and Data Analysis*, 25, 167-179.
- Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 129-132.
- Brownie, C., Boos, D. D., & Hughes-Oliver, J. (1990). Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls. *Biometrics*, 46, 259-266.
- Burnside, W. (1923). On errors of observation. *Proceedings of the Cambridge Philosophical Society*, 21, 482-487.
- Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, 7, 207-214.
- Conover, W. J., & Iman, R. I. (1982). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35, 124-129.
- Deshpande, J. V., Gore, A. P., & Shanubhogue, A. (1995). *Statistical analysis of nonnormal data*. NY: John Wiley & Sons.
- Diamond, W. J. (1981). *Practical experiment designs for engineers and scientists*. Belmont, CA: Lifetime Learning Publications, Wadsworth.
- Dissertation Abstracts Online*. (2000). <http://firstsearch.oclc.org>.
- Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34, 789-799.
- Fisher, R. A. (1937a). Editorial note. *Annals of Eugenics*, 7, 146-151.
- Fisher, R. A. (1937b). On a point raised by M. S. Bartlett on fiducial probability. *Annals of Eugenics*, 7, 370-375.
- Fisher, R. A. (1939a). The comparison of samples with possibly unequal variances. *Annals of Eugenics*, 9, 174-180.
- Fisher, R. A. (1939b). A note on fiducial inference. *The Annals of Mathematical Statistics*, 10, 383-388.
- Fisher, R. A. (1941). The asymptotic approach to Behrens's integral, with further tables for the d test of significance. *Annals of Eugenics*, 11, 141-172.
- Fisher, R. A., & Yates, F. (1957). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver & Boyd.

- Games, P. A., Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1*, 113-125.
- Gastwirth, J. L., & Podgor, M. J. (1992). Efficient robust rank tests for the location-scale problem. In Saleh, A. K. Md. E. (ed.) *Nonparametric statistics and related topics*. Amsterdam: Elsevier, p. 17-31.
- Grambsch, P., & O'Brien, P. (1991). The effects of transformations and preliminary tests for non-linearity in regression. *Statistics in Medicine, 10*, 697-709.
- Hájek, J., & Sidák, F. (1967). *Theory of rank tests*. Prague: Academic Press and Academia.
- Helmert, C. F. (1875). Über die Berechnung des wahrscheinlichen Fehlers aus einer endlichen Anzahl wahrer Beobachtungsfehler. *Zeitschrift für Mathematik und Physik, 20*, 300-303.
- Helmert, C. F. (1876). Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit in Zusammenhang stehende Fragen, *Zeitschrift für Mathematik und Physik, 21*, 192-219.
- Howell, J. F., & Games, P. A. (1974). The effects of variance heterogeneity on simultaneous multiple-comparison procedures with equal sample size. *British Journal of Mathematical and Statistical Psychology, 27*, 72-81.
- Huber, P. J. (1981). *Robust statistics*. NY: Wiley.
- Hyman, R. (1995). How to critique a published article. *Psychological Bulletin, 118*, 178-182.
- James, G. S. (1959). The Behrens-Fisher distribution and weighted means. *Journal of the Royal Statistical Society, 21*, 73-80.
- Jeffreys, H. (1937). On the relation between direct and inverse methods in statistics. *Royal Society of London Proceedings, Series A, 160*, 325-348.
- Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Annals of Eugenics, 10*, 48-51.
- Johnson, R. A., & Weerahandi, S. (1988). A Bayesian solution to the multivariate Behrens-Fisher problem. *Journal of the American Statistical Association, 83*, 145-149.
- Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). *The Functional Independence Measure: A new tool for rehabilitation*. In M. G. Eisenberg & R. C. Grzesiak (Eds.), *Advances in clinical rehabilitation* (Vol. 1). NY: Springer, p. 6-18.
- Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology, 30*, 213-221.
- Kohr, R. L. (1970). A comparison of procedures for testing  $\mu_1 = \mu_2$  with unequal n's and variances. Unpublished doctoral dissertation, The Pennsylvania State University.
- Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. *The Journal of Experimental Education, 43*, 61-69.
- Koschat, M. A., & Weerahandi, S. (1992). Chow-type tests under heteroscedasticity. *Journal of Business and Economic Statistics, 10*, 221-228.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association, 47*, 583-621.
- Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika, 58*, 213-217.
- Lepage, Y. (1975). Asymptotically optimum rank tests for contiguous location and scale alternatives. *Communications in Statistics, 4*, 671-687.
- Levene, H. (1960). Robust tests for equality of variance. In I. Olkin (Ed.) *Contributions to probability and statistics*. Palo Alto, CA: Stanford University Press, p. 278-292.
- Mann, H. B., & Whitney, d. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50-60.
- Maxwell, S. E., & Cole, D. A. (1995). Tips for writing (and reading) methodological articles. *Psychological Bulletin, 118*, 193-198.
- McCullough, R. S., Gurland, J., & Rosenberg, L. (1960). Small sample behaviour of certain tests of the hypothesis of equal means under variance heterogeneity. *Bimoetrika, 47*, 345-353.

- McSweeney, M., & Penfield, D. (1969). The normal scores test for the c-sample problem. *The British Journal of Mathematical and Statistical Psychology*, 20, 187-204.
- Mehta, J. S., & Srinivasa, R. (1970). On the Behrens-Fisher problem. *Biometrika*, 57, 649-655.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Nanna, M., & Sawilowsky, S. (1998). Analysis of Likert scale data in disability and medical rehabilitation evaluation. *Psychological Methods*, 3, 55-67.
- Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*. London: Unwin Hyman Ltd.
- Neyman, J., & Pearson, E. S. (1931). On the problem of k samples. *Bulletin internationale de l'Académie Polonaise des Sciences et des lettres (Cracovié), Sciences mathématiques, Série A*, 460-481.
- O'Brien, P. C. (1988). Comparing two samples: extension of the t, rank-sum, and log-rank tests. *Journal of the American Statistical Association*, 83, 52-61.
- Oshima, T. C., & Algina, J. (1992). Type I error rates for James's second-order test and Wilcoxon's  $H_m$  test under heteroscedasticity and non-normality. *British Journal of Mathematical and Statistical Psychology*, 45, 255-263.
- Pagurova, V. I. (1968). On comparison of mean values based on two normal samples. *Teoriya Veroyatnostei i Ee Primeneniya*, 13, 561-569.
- Podgor, M. J., & Gastwirth, J. L. (1994). On non-parametric and generalized tests for the two-sample problem with location and scale change alternatives. *Statistics in Medicine*, 14, 747-758.
- Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 59, 665-680.
- Ray, W. D., & Pitman, A. E. N. T. (1961). An exact distribution of the Fisher-Behrens-Welch statistic for testing the difference between the means of two normal populations with unknown variances. *Journal of the Royal Statistical Society*, 23, 377-384.
- SAS (1990). *SAS/STAT user's guide, Vol. 1*. (4<sup>th</sup> ed.) Cary, NC: SAS Institute.
- Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.
- Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, 60, 91-126.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the  $t$  test to departures from population normality. *Psychological Bulletin*, 111, 353-360.
- Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples  $t$  test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology*, 60, 240-243.
- Sawilowsky, S. S., Baerg, P., Boza, L. A. D., Kallmannsohn, M., Spencer, B., & Vollhardt, L. T. (April, 1991). Power analysis of the Brownie-Boos-Oliver t test for expected increases in treatment variability. Annual meeting of the American Educational Research Association, SIG/Educational Statisticians. Chicago, IL.
- Scheffé, H. (1943). On solutions of the Behrens-Fisher problem, based on the t-distribution. *Annals of Mathematical Statistics*, 14, 35-44.
- Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, 67, 93-101.
- SPSS (1993). *SPSS for Windows: Base system user's guide release 6.0*. Chicago: SPSS.
- Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, 5, 1055-1098.
- Student. (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- SYSTAT (1990). *SYSTAT: The system for statistics*. Evanston, IL: SYSTAT.
- Tan, W. Y., & Tabatabai, M. A. (1985). Some robust ANOVA procedures under heteroscedasticity and nonnormality. *Communications in Statistics*, 14, 1007-1026.
- Tiku, M. L., & Singh, M. (1981). Robust test for means when population variances are unequal. *Communications in Statistics*, 10, 2057-2071.

Tomarkin, A. J., & Serlin, R. c. (1986). Comparison of ANOVA under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90-99.

Troendle, J. F., Blair, R. C., Rumsey, D., & Moke, P. (1997). Parametric and non-parametric tests for the overall comparison of several treatments to a control when treatment is expected to increase variability. *Statistics in Medicine*, *16*, 2729-2739.

Wald, A. (1955). Testing for differences between the means of two normal populations with unknown standard deviations. *Selected papers in statistics and probability*. NY: McGraw-Hill.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-62.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*, 28-35.

Welch, B. L. (1949a). Further notes on Mrs. Aspin's tables. *Biometrika*, *36*, 243-246.

Welch, B. L. (1949b). Appendix to A. A. Aspin's tables. *Biometrika*, *36*, 293-296.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, *32*, 771-780.

Wiles, A. (January 15, 1996). J. Lynch (Ed.) Interview: Fermat's last theorem. BBC Horizon.

Yates, F. (1939). An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters. *Proceedings of the Cambridge Philosophical Society*, *35*, 579-591.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*, 165-170.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology*, *51*, 139-150.



## On The Misuse Of Confidence Intervals For Two Means In Testing For The Significance Of The Difference Between The Means

George W. Ryan      Steven D. Leadbetter  
Centers For Disease Control And Prevention

---

Comparing individual confidence intervals of two population means is an incorrect procedure for determining the statistical significance of the difference between the means. We show conditions where confidence intervals for the means from two independent samples overlap and the difference between the means is in fact significant.

Key words: Overlapping confidence intervals, significance tests, statistical tests of significance, tests for differences of means

---

### Introduction

When conducting a hypothesis test on the difference between two means (i.e.,  $H_0: \mu_1 - \mu_2 = 0$ ) or the special case of the difference between two proportions (i.e.,  $H_0: p_1 - p_2 = 0$ ) from two independent samples, some practitioners, researchers, and students may be tempted to compare the confidence intervals for the two individual means to determine the statistical significance of the difference. If the individual confidence intervals overlap, one might conclude, in error, that the means do not differ because of this overlap.

---

George W. Ryan is a Mathematical Statistician, Office of Statistics & Programming (OSP), National Center for Injury Prevention and Control (NCIPC), CDC, Atlanta, Georgia (e-mail: [gyr0@cdc.gov](mailto:gyr0@cdc.gov)). He is a graduate of Texas A&M University with over 20 years experience as a statistician in the federal government. Steven D. Leadbetter is a Mathematical Statistician, OSP, NCIPC, CDC, Atlanta, GA (e-mail: [SLeadbetter@cdc.gov](mailto:SLeadbetter@cdc.gov)). He received an M.S. in Applied Statistics from North Dakota State University and has more than 18 years experience as a statistician with the federal government. The authors thank Marcie-jo Kresnow and Scott R. Kegler for helpful comments.

We say that confidence intervals for means  $\mu_1$  and  $\mu_2$  computed from sample means  $\bar{x}_1$  and  $\bar{x}_2$ , where  $\bar{x}_1 \leq \bar{x}_2$ , overlap if the upper bound on  $\bar{x}_1$  exceeds the lower bound on  $\bar{x}_2$ . This misinterpretation of confidence intervals occurs widely in practice (Schenker & Gentleman, 2001); many researchers and even some statisticians mistakenly believe it. Accordingly, we consider the separate confidence intervals associated with the individual hypothesis tests for  $\mu_1$  and  $\mu_2$  (i.e.,  $H_{0_1}: \mu_1 = 0$  and  $H_{0_2}: \mu_2 = 0$ ) and the implications of attempting to test the hypothesis  $H_0: \mu_1 - \mu_2 = 0$  in terms of the individual confidence intervals associated with  $H_{0_1}$  and  $H_{0_2}$ .

Examples of overlapping confidence intervals for means that differ significantly are provided by Nelson (1989) and Barr (1969). Assuming a common known population variance, Nelson (1989) and Barr (1969) show that when given sample means from two normally distributed populations, the appropriate confidence interval for testing the hypothesis  $H_0: \mu_1 - \mu_2 = 0$  is based on the difference of the sample means,  $\bar{x}_1 - \bar{x}_2$ . We generalize this result to include the assumption of unequal sample variances and the special case of two proportions.

### Methodology

*Statistically Significant Difference of Two Means*  
Consider the case of independent random samples of size  $n_1$  and  $n_2$  from two populations with sample

means  $\bar{x}_1$  and  $\bar{x}_2$  and variances  $s_1^2, s_2^2$ . For simplicity, assume the population variances are equal and the populations are either normally distributed or the samples are sufficiently large so the assumptions of the Student's  $t$ -test are satisfied for the hypothesis tests and confidence intervals (Woodward, 1999). (This assumption will avoid any unnecessary complications with the distribution of the test statistic when the population variances are unequal.) The two sample means differ significantly at the .05 alpha level if the difference  $|\bar{x}_1 - \bar{x}_2|$  exceeds about 2 standard errors of the difference of the means (i.e.,  $|\bar{x}_1 - \bar{x}_2| \geq 2s_{\bar{x}_1 - \bar{x}_2}$ ).

For simplicity and clarity, because this discussion is in an instructional context, we use the quantity 2 as a sufficiently close approximation to the critical value of the Student's  $t$ -distribution at the .05 alpha level, which for large sample sizes will be close to the standard normal distribution critical value of 1.96. How can this difference hold if the individual confidence intervals for  $\mu_1$  and  $\mu_2$  overlap? If the confidence intervals overlap and the sample means  $\bar{x}_1$  and  $\bar{x}_2$  differ significantly, then (from Figure 1 below), it is necessary that  $s_{\bar{x}_1} + s_{\bar{x}_2} > s_{\bar{x}_1 - \bar{x}_2}$ . That is, the sum of the individual standard errors must exceed the standard error of the difference of the means.

An estimate of  $\sigma_{\bar{x}_1 - \bar{x}_2}^2$  is given by  $s_{\bar{x}_1 - \bar{x}_2}^2 = s^2(1/n_1 + 1/n_2)$ , where  $s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$  is an estimate of  $\sigma^2$  obtained by pooling  $s_1^2$  and  $s_2^2$  (Woodward, 1999). To be significant at the .05 alpha level, the difference in means  $|\bar{x}_1 - \bar{x}_2|$  must equal or exceed

$$2s \sqrt{1/n_1 + 1/n_2} \tag{1}$$

But for the confidence intervals to overlap, the difference between the means must be less than

$$2(s_1/\sqrt{n_1} + s_2/\sqrt{n_2}) \tag{2}$$

Accordingly, if  $|\bar{x}_1 - \bar{x}_2|$  is greater than or equal to (1) but less than (2), the difference of the means is significant and the individual confidence intervals overlap.

*Example.* The following data for two independent samples is taken from Woodward (1999). For the first sample,  $n_1 = 39, \bar{x}_1 = 6.168,$  and  $s_1 = 0.709$ ; for the second sample,  $n_2 = 11, \bar{x}_2 = 6.708,$  and  $s_2 = 0.803$ . The computed  $t$ -statistic for the test of the hypothesis  $H_0: \mu_1 - \mu_2 = 0$  is  $t(48) = -2.17$  (Woodward, 1999, p. 78) with a resulting p-value of .0351, indicating significance at the .05 alpha level. The 95% confidence intervals for  $\mu_1$  and  $\mu_2$  are (5.938, 6.398) and (6.169, 7.247), respectively. Accordingly, the sample means  $\bar{x}_1$  and  $\bar{x}_2$  differ significantly ( $p = .0351$ ) yet the confidence intervals overlap. Moreover, note the conditions from (1) and (2) above and in Figure 1 are satisfied; i.e.,  $2s_{\bar{x}_1} + 2s_{\bar{x}_2} > |\bar{x}_1 - \bar{x}_2| \geq 2s_{\bar{x}_1 - \bar{x}_2}$ ; for this example,  $.711 > .540 > .498$ .

*Statistically Significant Difference of Two Proportions*

Two independent proportions,  $p_1$  and  $p_2$ , may also be used to illustrate that overlapping confidence intervals do not imply nonsignificance of the observed difference. We now assume the samples are sufficiently large so that  $p_1$  and  $p_2$  (and hence their difference) are normally distributed. To be significant at the .05 alpha level, the difference  $|p_1 - p_2|$  in the proportions must equal or exceed

$$2 \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2} \tag{3}$$

However, individual confidence intervals for  $p_1$  and  $p_2$  will overlap if  $|p_1 - p_2|$  is less than

$$2(\sqrt{p_1(1-p_1)/n_1} + \sqrt{p_2(1-p_2)/n_2}) \tag{4}$$

using the quantity 2 as a sufficiently close approximation to the appropriate value (1.96) of the standard normal distribution. For  $0 < p_1, p_2 < 1,$  and  $n_1, n_2 > 1,$  the quantity (3) will always be less than (4). So, it could happen that  $|p_1 - p_2|$  is greater than or equal to (3) but less than (4), in which case the difference between the proportions would be significant and the confidence intervals would overlap.

Figure 1. Necessary conditions for overlapping 95% confidence intervals for two sample means differing significantly (using the quantity 2 as a sufficiently close approximation to the appropriate critical values of the Student's *t*-distribution).

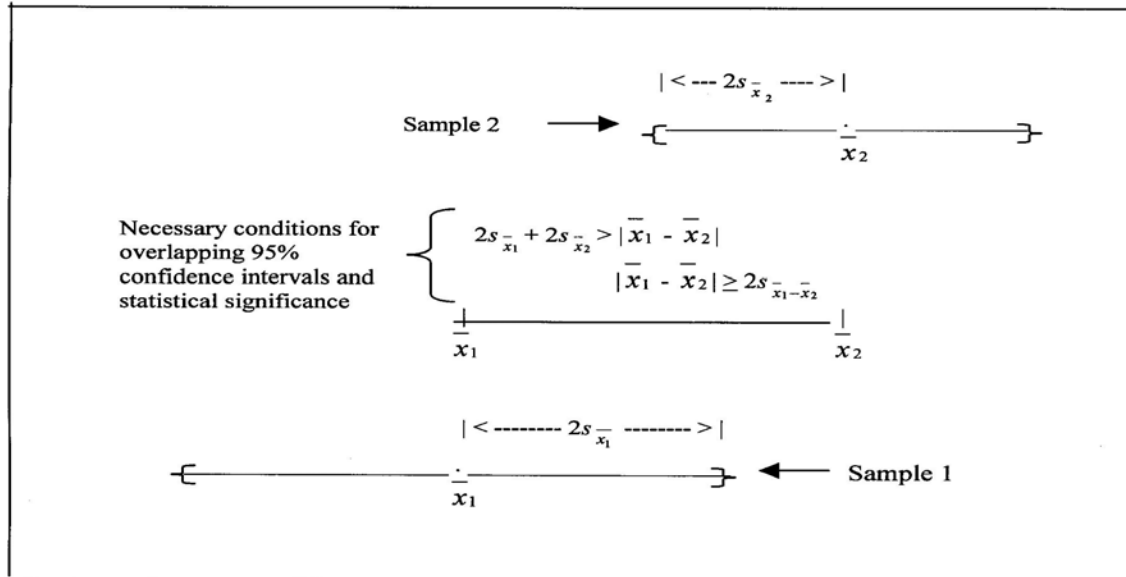
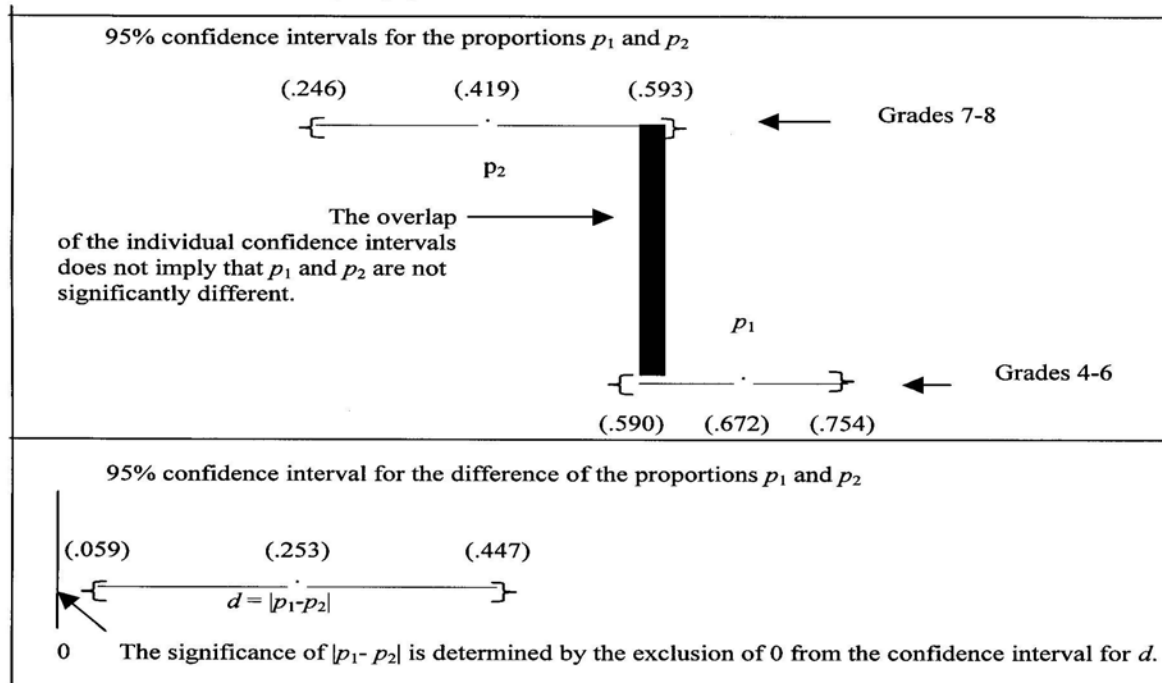


Figure 2. Texas Bicycle Helmet Study Data. Example: 95% confidence intervals for proportions of students agreeing ( $p_1$  in grades 4-6,  $p_2$  in grades 7-8) that "helmets should be worn" and the 95% confidence interval for  $d = |p_1 - p_2|$ .



## Results

The Texas Bicycle Helmet Study (Logan, Leadbetter, & Gibson, 1998) provides an example of two independent proportions  $p_1$  and  $p_2$  with overlapping confidence intervals and a significant difference between the proportions. Elementary and middle school students were surveyed over three time periods to assess their attitudes on such issues as helmet use, school rules, and social acceptability of bicycle helmets. In this example, let  $p_1$  be the proportion of students in grades 4 - 6 in survey period 3 who agree that students "must wear helmets" and  $p_2$  the corresponding proportion of students in grades 7 - 8 (see Figure 2 above). We are interested in testing  $H_0: p_1 = p_2$ . What result is obtained by observing the individual 95% confidence intervals? How does this result compare with the hypothesis test?

The upper bound of the confidence interval for  $p_2$  (.593) is greater than the lower bound for  $p_1$  (.590), leading some to conclude incorrectly that the observed difference  $p_1 - p_2$  is not significant. However, dividing the difference of the proportions (.253) by the standard error of the difference (.098) results in a test statistic of  $z = 2.58$ , which corresponds to a significance probability (p-value) of .0099. As shown previously, the individual confidence intervals overlap even though  $p_1$  and  $p_2$  differ significantly at the .05 alpha level provided  $|p_1 - p_2|$  is less than twice the sum of the individual standard errors of  $p_1$  and  $p_2$ . In this example,  $p_1$  and  $p_2$  differ significantly, but the individual confidence intervals overlap as the difference  $p_1 - p_2$  (.253) is less than twice the sum of the individual standard errors ( $2(.042 + .089) = .262$ ).

Of course, the proper interpretation of hypothesis testing in the context of confidence intervals consists (using the present example) of the estimated difference  $d = p_1 - p_2$  with its associated lower and upper bounds to see if *that* confidence interval includes zero (see Figure 2) (Woodward, 1999). For any significance level, failure of the associated confidence interval to "cover" zero will always indicate significance in the corresponding hypothesis test. To correctly interpret the relationship between confidence intervals and hypothesis tests, one needs to use the confidence interval of the difference.

## Conclusion

Our purpose has been to show that an overlap of individual confidence intervals for two means or proportions does not necessarily indicate that the difference between the means is nonsignificant. The proper interpretation of confidence intervals is important because of their increased use in recent years as an inferential tool in preference to traditional hypothesis testing (Chow, 1996). In disciplines such as medicine (Gardner & Altman, 1986), epidemiology (Savitz, Tolo, & Poole, 1994), education (Nix & Barnette, 1998), and psychology (Krantz, 1999), many believe that confidence intervals are more meaningful and easier to interpret than tests of significance.

This erroneous use of individual confidence intervals to determine the significance of the difference between two means could lead one to fail to reject the hypothesis of no difference when the difference is indeed significant. This misuse of individual confidence intervals results in an overly conservative test (Schenker & Gentleman, 2001). In the Texas Bicycle Helmet Study, which used .05 as the stated alpha level, the actual significance probability (p-value) was .0099, indicating a significant difference of means.

The erroneous interpretation of overlapping confidence intervals would lead one to conclude otherwise. The potential for misinterpretation is even more profound if the observations are taken from a sample of paired data since the standard error of the difference (between the observations in each pair) can be considerably smaller (assuming the sample means are positively correlated) than the standard errors of the means from the individual samples (Woodward, 1999). Using the individual confidence intervals here to test the hypothesis  $H_0: d = 0$  ( $d$  being the difference within each paired observation) would be an exceedingly conservative procedure.

To indicate how individual 95% confidence intervals can overlap even when the means differ significantly, we generated confidence intervals for two proportions  $p_1$  and  $p_2$  for a range of sample sizes. Using values of  $p_1 = .65$  and  $p_2 = .40$  (chosen because they are comparable to the values in the previous example) and, for simplicity, equal size samples from each

population (i.e.,  $n_1 = n_2 = n$ ), we computed confidence intervals for  $p_1$  and  $p_2$ . Percent overlap is defined as the ratio of the amount of overlap of the confidence intervals to the difference  $p_1 - p_2$ . For sample sizes ranging from 30 to 57 from each population, the individual confidence intervals overlap and the two proportions differ significantly (see Figure 3).

For  $n < 30$ , the individual confidence intervals overlap, but the difference of the proportions is no longer significant at the .05 alpha level. For  $n > 57$ , the proportions are significantly different, but the confidence intervals no longer overlap. It is within the range of sample sizes from 30 to 57 (for the selected values of  $p_1$  and  $p_2$ ) that one could erroneously conclude that the difference  $p_1 - p_2$  is significant on the basis of overlapping confidence intervals. As the percent overlap decreases, so too does the significance probability (see Figure 3). Accordingly, the consequences of misinterpretation are greater as the overlap becomes smaller. In the example in Figure 2, the percent overlap is  $(.593 - .590) / (.672 - .419)$ , or 1.2%, but the significance probability, as previously noted, is .0099.

Note that for any value  $n$  selected within the range (30, 57) in Figure 3 (next page) for equal sample sizes ( $n_1 = n_2 = n$ ), the difference  $p_1 - p_2$  (.25) will be greater than expression (3) and less than (4), the conditions previously noted for overlapping 95% confidence intervals for two significantly different proportions.

Why does this problem persist? Some users may be accustomed to viewing graphical and other displays of data, such as results of multiple range tests, in which overlapping segments of output do indicate nonsignificant differences. They may jump to the erroneous conclusion that overlapping confidence intervals imply that the difference of the means is nonsignificant. Another notion that may contribute to the belief that overlapping confidence intervals imply a nonsignificant difference is the case of nonoverlapping confidence intervals for proportions from two independent samples (Centers for Disease Control and Prevention, 1995).

In the case of two proportions, from the conditions noted in (3) and (4), the sum of the individual standard errors always exceeds the standard error of the difference. It then follows

that if the confidence intervals do not overlap, the difference of the proportions is indeed significant. This fact may lead some to conclude that two proportions do not differ significantly if their confidence intervals do overlap.

So what do the individual confidence intervals say about the difference between the means? These intervals are statements only about the variability of each individual estimate; they say nothing about their difference. To determine the significance of the difference in the context of a confidence interval, lower and upper bounds for the difference can be computed quite routinely once the standard error of the difference between the means has been obtained. Only by looking at the lower and upper confidence limits for this difference (see Figure 2) and noting whether the interval includes (or excludes) zero, can one determine the statistical significance of the difference.

#### References

- Barr, D. R. (1969). Using confidence intervals to test hypotheses. *Journal of Quality Technology, 1*, 256-258.
- Centers for Disease Control and Prevention. (1995). *Healthy people 2000 statistical notes*. Atlanta, GA.
- Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks, CA: Sage Publications.
- Gardner, M. J. & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Statistics in Medicine, 292*, 746-750.
- Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44*, 1372-1381.
- Logan, P., Leadbetter, S., & Gibson, R. E. (1998). Evaluation of a bicycle helmet giveaway program – Texas, 1995. *Pediatrics, 101*, 578-582.
- Nelson, L. S. (1989). Evaluating overlapping confidence intervals. *Journal of Quality Technology, 21*, 140-141.

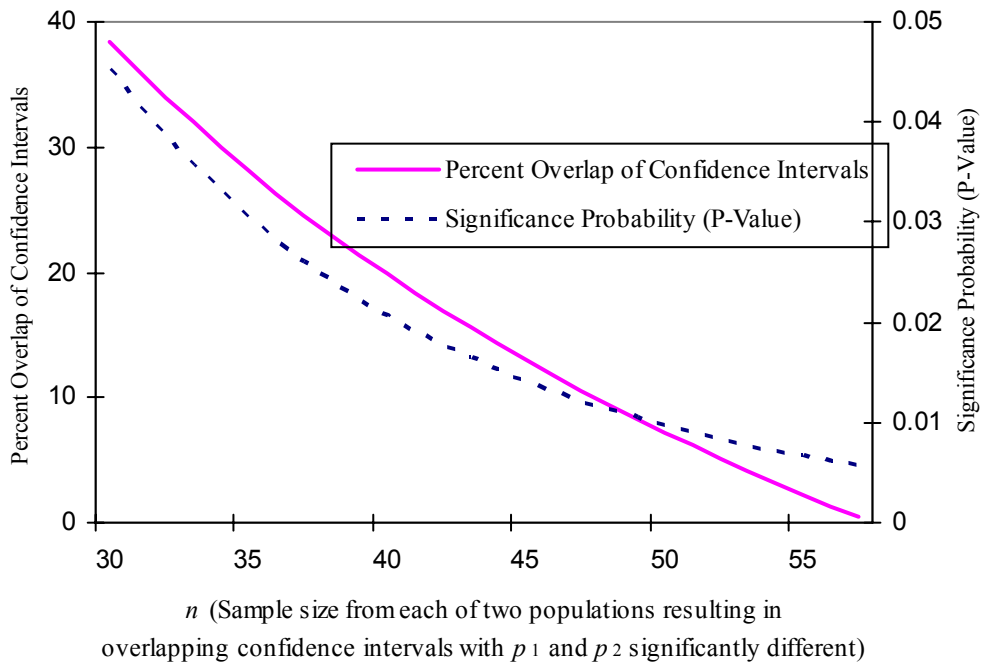
Nix, T. W. & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5, 3-14.

Savitz, D. A., Tolo, K-A., & Poole, C. (1994). Statistical significance testing in the American Journal of Epidemiology. *American Journal of Epidemiology*, 139, 1047-1052.

Schenker, N. & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55, 182-186.

Woodward, M. (1999). *Epidemiology: Study design and data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Figure 3. Percent overlap of confidence intervals for  $p_1$  and  $p_2$  and significance probabilities ( $30 \leq n \leq 57, p_1 = .65, p_2 = .40$ ).



## *Early Scholars* Best Regression Model Using Information Criteria

Phill Gagné  
Department of Measurement  
Statistics and Evaluation  
University of Maryland – College Park

C. Mitchell Dayton  
Department of Measurement  
Statistics and Evaluation  
University of Maryland – College Park

---

The accuracy of AIC and BIC is evaluated under simulated multiple regression conditions, varying number of total and valid predictors,  $R^2$ , and  $n$ . AIC and BIC were increasingly accurate as  $n$  increased and as total predictors decreased. Interactions of the ratio of valid/total predictors affected accuracy.

Key words: AIC, BIC, simulated regression, information criteria

---

### Introduction

Exploratory model building is often used within the context of multiple regression (MR) analysis. As noted by Draper and Smith (1998), these undertakings are usually motivated by the contradictory goals of maximizing predictive efficiency and minimizing data collection/monitoring costs. A popular compromise has been to adopt some strategy for selecting a “best” subset of predictors.

Many different definitions of best can be found in the literature, including incremental procedures such as forward selection MR, backward elimination MR, stepwise MR, all-possible subsets MR with criteria related to residual variance, multiple correlation, Mallows  $C_p$ , etc. Incremental procedures are efficient, computationally, but do not necessarily result in

the selection of an unconditionally best model. For example, as usually implemented, forward selection MR includes additional variables in the regression model based on maximizing the increment to R-squared from step to step. At the third step, for example, the model contains the best three predictors only in a conditional sense. Also, the modifications to forward selection incorporated into stepwise MR do not guarantee finding the best three predictors.

In contrast to incremental procedures, all-possible subsets does choose a best model for a fixed number of predictors but not necessarily an overall best model. For the  $m^{\text{th}}$  model based on  $p_m$  out of a total of  $p$  independent parameters, Mallows  $C_p$ , for example, utilizes a criterion of the form  $SS_m / \hat{\sigma}_e^2 - [n - 2(p_m + 1)]$  where  $\hat{\sigma}_e^2$  is the residual variance estimate based on the full model (i. e., the model with all  $p$  predictors). Models with values close to  $p_m + 1$  are best in a final prediction error (FPE) sense. Thus, a best model can be identified for fixed values of  $p_m$ , but there is no general method for selecting an overall best model.

Akaike (1973) adopted the Kullback-Leibler definition of information,  $I(f; g)$ , as a natural measure of discrepancy, or asymmetrical distance, between a true model,  $f(y)$ , and a proposed model,  $g(y|\beta)$ , where  $\beta$  is a vector of parameters. Based on large-sample theory, Akaike derived an estimator for  $I(f; g)$  of the form:

---

Phill Gagné is a doctoral student in the Department of Measurement, Statistics and Evaluation at the University of Maryland – College Park. E-mail: blueplanet88@hotmail.com. His research interests include factor analysis, SEM, and personality assessment. C. Mitchell Dayton is Professor of Measurement, Statistics and Evaluation at the University of Maryland, College Park, Maryland. E-Mail: cd4@umail.umd.edu. His current research interests are focused on the areas of latent variable analysis and model comparison procedures.

$$AIC_m = -2Ln(L_m) + 2 \cdot k_m \quad ,$$

where  $L_m$  is the sample log-likelihood for the  $m^{\text{th}}$  of  $M$  alternative models and  $k_m$  is the number of independent parameters estimated for the  $m^{\text{th}}$  model. The term,  $2 \cdot k_m$ , may be viewed as a penalty for over-parameterization. The derivation of AIC involves the notion of loss of information that results from replacing the true parametric values for a model by their maximum likelihood estimates (MLE's) from a sample. In addition, Akaike (1978b) has provided a Bayesian interpretation of AIC.

A min(AIC) strategy is used for selecting among two or more competing models. In a general sense, the model for which  $AIC_m$  is smallest represents the "best" approximation to the true model. That is, it is the model with the smallest expected loss of information when MLE's replace true parametric values in the model. In practice, the model satisfying the min(AIC) criterion may or may not be (and probably is not) the "true" model since there is no way of knowing whether the "true" model is included among those being compared. Unlike traditional hypothesis testing procedures, the min(AIC) model selection approach is holistic rather than piecemeal. Thus, for example, in comparing four hierarchic linear regression models, AIC is computed for each model and the min(AIC) criterion is applied to select the single "best" model. This contrasts with the typical procedure of testing the significance between models at consecutive levels of complexity. An excellent and more complete introduction to model selection procedures based on information criteria is presented by Burnham and Anderson (1998).

Typically, for regression models, the number of independent parameters,  $k_m$ , is equal to the number of predictor variables in the equation plus two since, in addition to partial slope coefficients, an intercept and residual variance term are estimated. It should be noted that the maximum likelihood estimator for the residual variance is biased (i. e., the denominator is the sample size,  $n$ , rather than  $n - p_m - 1$  for a  $p_m$ -predictor model). In particular, for  $p$  predictors based on a normal regression model (i. e., residuals assumed to be normally distributed with homogeneous variance), the log(likelihood) for the

model is:  $-5n \cdot (\ln(2\pi) + \ln(SS_e / n) + 1)$  where  $SS_e$  is the sum of squared residuals. Then, the Akaike information measure is:

$$AIC = n(\ln(2\pi) + \ln(SS_e / n) + 1) + 2(p_m + 2).$$

The Akaike model selection procedure entails calculating AIC for each model under consideration and selecting the model with the minimum value of AIC as the preferred, or "best," model. In the context of selecting among regression models, a "best" model can be selected for each different size subset of predictors as well as overall.

AIC, which does not directly involve the sample size,  $n$ , has been criticized as lacking properties of consistency (e.g., Bozdogan, 1987; but see Akaike, 1978a for counter arguments). A popular alternative to AIC presented by Schwarz (1978) and Akaike (1978b) that does incorporate sample size is BIC where:

$$BIC_m = -2Ln(L_m) + \ln(n) \cdot k_m.$$

BIC has a Bayesian interpretation since it may be viewed as an approximation to the posterior odds ratio. Note that BIC entails heavier penalties per parameter than does AIC when the sample size is eight or larger. When the order of the model is known and for reasonable sample sizes, there is a tendency for AIC to select models that are too complex and for BIC to select models that are too simple. In fact, the relative tendencies for the occurrence of each type of misspecification can be derived mathematically as shown by McQuarrie and Tsai (1998). The tendency for AIC to select overly complex models in cases where complexity is known has been interpreted as a shortcoming of this measure. Hurvich and Tsai (1991), for example, argue for a modified version of AIC that incorporates sample size. In practical applications, however, the performance of criteria such as AIC and BIC can be quite complex.

AIC was originally developed by Akaike within the context of relatively complex autoregressive time series models for which he presented some simulation results (Akaike, 1974). Bozdogan (1987) compared rates of successful model identifications for AIC and CAIC (a close kin of BIC) for a single cubic model with various



error structures. Hurvich and Tsai (1991) compared AIC and their own consistent estimator, AICC, for a normal regression case and for a complex time series. Bai et al. (1992) compared AIC and several modifications of AIC within the context of multinomial logistic regression models. Although each of these previous studies has investigated the use of AIC and related criteria in exploratory frameworks, the present study expands the focus to applications of multiple regression analysis that are more typical of a behavioral science setting. More specifically, AIC and BIC were investigated under a variety of realistic scenarios.

### Methodology

AIC and BIC were evaluated under several simulated multiple regression conditions. Data were collected regarding the accuracy of both information criteria for each condition and the nature of the incorrect choices. The accuracy of an information criterion was defined as the percentage of iterations in which it selected the correct model. Incorrect model selections fell into one of three categories: 1) Low: The chosen model had too few predictors in it; 2) High: The chosen model had too many predictors in it; 3) Off: The chosen model had the correct number of predictors but included one or more that had a correlation of 0 with the criterion without including one or more that had a nonzero correlation with the criterion.

The number of total predictors, the number of valid predictors, R-squared, and sample size were manipulated. For total number of predictors,  $p$ , the values of 4, 7, and 10 were chosen. These values are a reasonable representation of the number of predictors found in applied research settings and they are sufficiently different to illustrate potential relationships between  $p$  and accuracy of the information criteria. With 4 total predictors, conditions with 2, 3, and 4 valid predictors ( $v$ ) were simulated; with 7 total predictors, conditions with 2 through 7 valid predictors were simulated; and with 10 total predictors, conditions with 2 through 8 valid predictors were simulated. For  $p = 10$ , 9 and 10 valid predictors were not included because predictor-criterion correlations for a ninth and tenth valid predictor at  $R^2 = .1$ , after controlling for the first eight predictors would have been trivially

small. Furthermore, research contexts rarely incorporate 9 or 10 valid predictors for a single criterion.

Three values of R-squared, .1, .4, and .7, were evaluated. These values were chosen to represent small, moderate, and large multiple correlations, respectively. They were also chosen to allow for consideration of accuracy trends that were a linear function of R-squared.

Each combination of the above factors was tested with sample sizes that were 5, 10, 20, 30, 40, 60 and 100 times the number of total predictors. Relative sample sizes were used rather than absolute sample sizes, because sample size recommendations in multiple regression are typically a function of the number of predictors in the model. These values for relative sample size were chosen to simulate conditions that were below generally accepted levels, at or somewhat above generally accepted levels, and clearly above generally accepted sample sizes.

All simulations were carried out by programs written and executed using SAS 8.0, and 1000 iterations were conducted for each condition. The simulated data were generated for each condition based on a correlation matrix with the designated number of nonzero correlations between predictors and the criterion. The correlations in each combination increased from zero in a linear fashion based on their squared values, such that the  $r^2$ -values summed to the designated  $R^2$ -value. All correlations among predictors were set at 0. Although, in applied work, predictors are not independent of each other, this design does not lose generalizability since this is equivalent to residualizing the predictor-criterion correlations for all but the strongest predictor to compute R-squared, which results in all these intercorrelations becoming 0, regardless of their original values.

### Results

#### Best Overall Models

The valid predictor ratio,  $VPR = v/p$ , is defined as the ratio of valid predictors ( $v$ ) to total predictors ( $p$ ). For purposes of interpreting accuracy in selecting true models, values of at least 70% were considered satisfactory. The percentage of correct selection is presented for AIC and BIC in Tables 1 and 2 (see Appendix A). Results based on sample size sorted by total numbers of variables

equal to 4, 7 and 10 are summarized as graphs in Figures 1, 2, and 3, respectively (shown following tables in Appendix A).

### BIC

The accuracy of BIC for selecting the best overall model consistently improved as sample size increased and as R-squared increased. In general, accuracy declined with increases in the total number of predictors,  $p$ , with an exception being the behavior for two valid predictors, where accuracy steadily improved as  $p$  increased. The relationship of accuracy to VPR was not as straightforward, being complicated by interactions with sample size, R-squared, and  $p$ . For all combinations of R-squared and total number of predictors, there was an inverse relationship between accuracy and VPR for values of  $p$  at  $n = 5p$ . For  $R^2 = .1$ , this relationship held across all sample sizes, with the differences between VPR's generally increasing with sample size. For  $R^2 = .4$ , the differences in accuracy between the VPR's within  $p$  slowly decreased, with the mid-range VPR's consistently being superior to the others at the two largest relative sample sizes. For  $R^2 = .7$ , there was an inverse relationship between VPR and accuracy at the lowest sample sizes; the relationship became direct, however, by  $n = 30p$  with  $p = 7$ , and  $n = 20p$  at 4 and 10 total predictors.

For  $R^2 = .1$ , the accuracy of BIC was generally low. In only 10 of the 112 combinations in the simulation design did BIC achieve acceptable accuracy, doing so when  $n \geq 400$  with two valid predictors,  $n \geq 600$  with three valid predictors, and at  $n = 1000$  with a VPR of 4/10. For  $R^2 = .4$ , the accuracy of BIC improved. For  $v = 2$ , sample sizes of  $10p$  were adequate to achieve acceptable accuracy. As VPR increased within  $p$ , and as  $p$  increased, the sample size necessary for acceptable accuracy also increased. At VPR's of 7/7 and 8/10, for example, acceptable accuracy was not achieved until  $n = 60p$ , while at VPR = 4/4, BIC was 69.2% accurate at  $n = 30p$  and 80.5% accurate at  $40p$ .

For  $R^2 = .7$ , BIC was quite accurate at all but the smallest relative sample size. At  $n = 5p$ , BIC's accuracy was only acceptable with VPR = 2/4. At  $n = 10p$ , only VPR's of 7/7, 7/10, and 8/10 failed to achieve acceptable accuracy. For the remaining

relative sample sizes with  $R^2 = .7$ , BIC was at least 80% accurate.

### AIC

Like BIC, the accuracy of AIC at selecting the best overall model consistently declined as the total number of predictors was increased. This was the only similarity in the pattern of results for AIC and BIC. The change in accuracy of AIC was not stable across any other single variable.

AIC was consistently at its worst at the smallest sample sizes, with improved accuracy attained with medium sample sizes. For larger sample sizes, AIC behaved nearly at its asymptote, although rarely at or near 100% accuracy. Only VPR's of 4/4 and 7/7 approached 100% accuracy, doing so at the higher relative sample sizes with  $R^2 = .4$ , and doing so for  $n \geq 30p$  with  $R^2 = .7$ . As R-squared increased, each VPR behaved asymptotically at gradually smaller relative sample sizes. Lower VPR's stabilized around their asymptotes sooner, in terms of sample size, than higher VPR's due to a general tendency for the higher VPR's to be less accurate at the smaller sample sizes and due to the fact that higher VPR's consistently had higher asymptotes.

For the combinations with  $R^2 = .1$ , AIC achieved acceptable levels of accuracy even less frequently than did BIC, breaking the 70% barrier in only two cases:  $n = 400$  at VPR's of 2/4 and 3/4. With  $R^2 = .4$ , AIC did poorly for  $p = 10$  with only the  $v = 8$ ,  $n = 1000$  case reaching satisfactory accuracy. At VPR = 7/7, AIC performed well for sample sizes of at least  $30p$ .

AIC achieved acceptable accuracy at VPR's of 2/4, 3/4, and 4/4 by  $n = 20p$  (albeit asymptotically for 2/4). For  $R^2 = .7$ , all VPR's with  $p = 4$ , reached acceptable accuracy by  $10p$  (again asymptotically for 2/4). With VPR = 5/7, the accuracy of AIC again appeared asymptotic at 70%, but the VPR's 6/7 and 7/7 demonstrated acceptable accuracy for all but the smallest sample size. With eight valid predictors out of 10 total, AIC's accuracy seemed to be asymptotic for a value just above 70% at  $n \geq 30p$ .

### Comparison of BIC and AIC

At VPR's of 4/4 and 7/7, AIC was consistently as good as or better than BIC at selecting the correct overall model regardless of sample size and R-squared. With  $R^2 = .1$ , AIC

outperformed BIC at all sample sizes when the  $VPR > .5$ . For  $R^2 = .4$ , AIC consistently outperformed BIC only at  $n = 5p$  and  $n = 10p$  in conjunction with  $VPR$ 's above  $.5$ . For  $R^2 = .7$  and  $VPR > .5$ , AIC outperformed BIC only at  $n = 5p$  and for all other cases BIC outperformed AIC.

#### Patterns of Misselection

Unlike the accuracy patterns of BIC and AIC, patterns of incorrect choices are nearly identical and relatively straightforward. The incorrect decisions made by both AIC and BIC tended to be in the direction of more complex models when sample size was large and valid predictor ratio was low. At lower sample sizes and higher valid predictor ratios, both criteria tended to select models that were too simple.

The rates of change from errors of complexity to errors of simplicity, however, were appreciably different for AIC and BIC. As sample size increased with decreasing  $VPR$ , incorrect decisions by BIC tended toward simpler models until reaching the higher relative sample sizes with the lower  $VPR$ 's. AIC, by contrast, made more errors of simplicity than of complexity only at the combination of the lower sample sizes and higher  $VPR$ 's.

Results were also obtained for incorrectly selecting models with the correct number of predictors but not the actual best predictors. This type of error occurred more often with AIC than with BIC and in general, it happened more often at smaller sample sizes, smaller  $R^2$ -values, and for more total predictors. The relationship between  $VPR$ 's and the frequency of this type of incorrect selection interacted with  $R$ -squared and sample size. For  $R^2 = .1$ , these errors occurred predominantly at lower relative sample sizes with lower  $VPR$ 's. As  $VPR$  increased, the distribution became slightly quadratic, with the error occurring most at the moderate sample sizes and tapering to either side of the middle. At the higher values of  $VPR$ , the larger relative sample sizes contained the highest frequencies of this type of error.

For  $R^2 = .4$ , incorrectly selecting the right number but wrong set of predictors was generally limited to the lower sample sizes with the overall frequency dropping off rapidly after  $VPR = .5$ . For  $R^2 = .7$ , this type of error was rare; at no sample size above  $5p$  was the frequency greater than 4.3% of the iterations, the frequency never exceeded

10% for BIC and only at  $VPR$ 's of 7/10 (.136) and 8/10 (.139) did it exceed 10% for AIC.

#### Conclusion

The results of the present study suggest that different multiple regression scenarios in applied research call for different information criteria for selecting the best set of predictors. As is so often the recommendation in research, the larger the sample sizes the better; both BIC and AIC were increasingly more accurate as sample size increased. The information criteria were also generally more accurate as the number of total predictors decreased, although the reverse was true of BIC with two valid predictors. The results also provide some unfortunately complex recommendations for accuracy based on interactions of  $VPR$  with other facets of model conditions.

When all, or nearly all, predictors in a set are valid predictors, AIC is as good as or better than BIC at selecting the best overall model at every sample size and  $R^2$ -value tested. When  $R$ -squared is low, the advantage of AIC at higher valid predictor ratios is essentially moot, because at higher  $VPR$ 's neither information criterion reached satisfactory accuracy (except AIC at  $VPR = 3/4$  and  $n = 100p$ ). With higher multiple correlations, however, AIC was at least 70% accurate at high  $VPR$ 's and sample sizes of 20 to 30 times the number of predictors (with a negative relationship between sample size and  $R$ -squared required for good accuracy). For  $VPR$ 's above  $.5$  but below  $.8$ , sample size affects the relative performance of BIC and AIC. AIC is the better choice for relative sample sizes below  $30p$  when  $R^2 < .7$ . BIC is generally the better choice for relative sample sizes of at least  $30p$  or when  $R^2 \geq .7$ , with one exception in the current study at  $VPR = 3/4$  and  $R^2 = .1$  in which AIC is better across sample size. It should be noted, however, that with  $VPR$ 's in the  $.5$  to  $.8$  range and relative sample sizes below  $30p$ , neither AIC nor BIC reached satisfactory accuracy with  $R^2 < .7$ , so AIC's advantage in such situations may not have practical importance.

For  $VPR$ 's  $\leq .5$ , BIC performed uniformly better than AIC. The importance of this advantage was related to  $R$ -squared. With small multiple correlations, BIC only achieved satisfactory

accuracy at low VPR's for relatively large sample sizes ( $n \geq 400$ ). At moderate levels of R-squared, BIC begins to perform well at lower relative sample sizes (20p with 3 valid predictors and 10p at  $v = 2$ , with  $R^2 = .4$ ) when the VPR is low. At extremely high values of R-squared, BIC is at least 70% accurate with sample sizes that are 10 times the number of predictors when VPR is low.

The sample sizes chosen for the present study seemed to provide a reasonable illustration of the patterns of accuracy at fixed relative sample sizes. There were, however, very few conclusions that could be made based on absolute sample size. Restructuring the tables and charts to line up sample sizes would line up only similar sample sizes, the conclusions of which would be confounded by having only similar valid predictor ratios. It might therefore be fruitful to investigate patterns of the accuracy of information criteria as a function of absolute sample size.

#### References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723.

Akaike, H. (1978a). A new look at the Bayes procedure. *Biometrika*, 65, 53-59.

Akaike, H. (1978b). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, 30, 9-14.

Bai, Z. D., Krishnaiah, P. R., Sambamoorthi, N., & Zhao, L. C. (1992) Model selection for log-linear model. *Sankhya B*, 54, 200-219.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

Burnham, K. P. & Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Draper, N. R. & Smith, H. (1998) *Applied Regression Analysis* (3<sup>rd</sup> ed.). New York: Wiley.

Hurvich, C. M. & Tsai, C-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, 78, 499-509.

McQuarrie, A. D. R. & Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Appendix A: Tables & Figures

Table 1. Percentage of correct model selection for AIC.

R <sup>2</sup>	p	4					7					10					
		2	3	4	2	3	4	5	6	7	2	3	4	5	6	7	8
.1	n=5p	9.7	3.9	1.8	8.8	2.2	0.7	0.4	0.1	0	8.7	3.3	0.4	0.2	0.1	0.1	0
	10p	17.2	4.8	2.1	19	6	1.9	0.5	0.1	0.1	14.5	8.6	2.8	1.6	0.5	0	0
	20p	35.5	18.2	5.7	29.3	20.2	10.5	5.6	2	1.3	20.9	18.1	10.6	6.5	3.4	1.7	0.2
	30p	48	30.5	13.3	38	33	20.5	11	4.6	2.7	23	21.6	19.2	13.5	7.8	4.6	2.3
	40p	57.8	42.1	23	37.6	39.8	30.3	20.3	12	6.1	23.6	26.5	23.6	20.3	13.2	8.3	7
	60p	66.6	59.5	39.6	43.5	43.3	42.4	37.7	25.9	19.7	27.1	28.6	28.8	30.6	25.7	20.1	16.7
	100p	72.7	73.6	69.3	40.1	47.7	53.1	53.7	51.1	40.3	25.1	29.1	34.3	37	40.1	39.4	35.8
.4	n=5p	44.4	25.3	13.3	28.5	21.9	13.4	9.7	6	2.5	16.6	19.6	13.8	8.1	6.6	3.9	2.2
	10p	61.1	54	39.4	35.9	41.4	40.2	32.6	24.2	16.9	22.9	25.4	29.1	25.9	21.9	19.5	15.6
	20p	68.7	77.2	74.7	43	46.5	55.3	59.2	56.6	49.2	24.3	31.8	33.1	39.1	39.4	41.6	40.5
	30p	69.3	81.5	91.8	39.9	49.1	59.7	66.2	70.4	72.7	23.5	29.7	37.7	38.9	48.3	49.9	56.1
	40p	70.1	82.6	95.9	43.9	48.4	57.1	67	79.9	88.2	25	28.7	33.7	39.6	47	54.6	64.8
	60p	71.4	84.2	99.2	39.7	49.7	61.1	71.7	83.4	95.8	22	28.4	33.9	39.3	48.1	60.1	67.3
	100p	69.6	82.2	100	43.4	48	59.7	68.3	85.5	99.5	24.5	32.2	35	44.5	48.5	61.7	70.6
.7	n=5p	60	62.4	61.8	35	41.6	41.5	45	45.4	38.7	19.1	26	29.5	27.5	33.1	31.3	28.2
	10p	69.3	77.7	92.5	38.2	48.1	56.8	66.2	72.9	76.8	22.8	26.2	31.3	39.1	42.9	50.1	54.3
	20p	67	81.9	100	41.6	52	58.6	66.8	84.4	95.4	23.5	29.7	33.2	40.7	47.6	56.9	68.3
	30p	69.2	84.5	100	40.3	49.3	59.5	68.8	84.8	99.7	22.1	30.8	35.3	40.2	49.2	61.2	70.3
	40p	71	83.6	100	39.9	47.7	61.3	68.6	83.1	100	27.4	28.8	33.7	40.6	49.6	60.6	72.6
	60p	70.9	83.2	100	39.2	48.6	59.3	71.3	84.5	100	26.7	32.2	35.1	42.2	50	59.9	70.5
	100p	71.8	83.8	100	44.2	48	61.3	70.7	82.2	100	25.7	30.9	35.6	43.2	49.2	58.3	73.4

Table 2. Percentage of correct model selection for BIC.

R <sup>2</sup>	p	4					7					10					
		2	3	4	2	3	4	5	6	7	2	3	4	5	6	7	8
0.1	n=5p	6	2	0.3	7.7	0.9	0.1	0.1	0	0	8.2	1.4	0.1	0	0	0	0
	10p	8.8	1.5	0	16.3	1.5	0.1	0	0	0	21.5	3.9	0.1	0.2	0	0	0
	20p	22.3	4.1	0.3	34.6	8.5	1.4	0.1	0	0.1	48.5	17.6	3.8	0.1	0.1	0	0
	30p	34.3	7.5	0.9	57.9	21.3	4.2	0.5	0	0	69	32.4	11.6	2.2	0.4	0.3	0
	40p	45.8	12.4	1.7	69	29.2	13.3	1.9	0.4	0.1	82.9	46.8	19.5	5.6	1.3	0.2	0
	60p	68.3	26.8	7.2	86.8	53.6	25.1	6.8	1.2	0.3	89.5	71.2	40.5	18	4.9	1.4	0.7
	100p	88.2	53.2	21.1	94.2	79.8	49.9	24.3	8.9	2.9	91.8	91	70.4	46.9	23.5	10.4	3.4
0.4	n=5p	44.5	17.9	6.5	48.4	23.5	8.3	4.1	1.4	0.2	50.4	35.5	15.2	5.7	2.1	0.2	0.2
	10p	72.2	40.1	17.7	73.1	58.8	33.8	17.5	5.7	2.7	74.8	63.4	48.3	27.3	11.7	5.3	2.2
	20p	88.8	74.1	46.4	86.9	85.9	69.8	46.2	29.1	13.6	83.1	86	74.9	63.5	44.7	28.2	14.9
	30p	93.1	87.9	69.2	89.8	90.4	84.3	71.9	51.2	33.1	85.9	87.4	88.8	81.4	66.7	52.4	36
	40p	94.7	94.1	80.5	91.7	92.9	90.5	82.2	67.3	50.7	88.5	90	90.5	89.3	79.1	65.9	52.5
	60p	95.5	97.2	93.3	91.8	94.6	96.7	92.3	87.3	70.6	90.6	91.1	94.3	94.4	92.1	86.2	75.6
	100p	97	98.6	99.3	94.6	94.1	97.1	97.4	96.7	91.9	92.2	94.6	94.3	95.8	96.8	97.3	93.4
0.7	n=5p	72.7	62.8	45.7	65.4	67.2	54.1	43.7	33.1	17.3	58	61.6	61.3	50.6	38.5	26.8	18.2
	10p	88.2	87.4	80.6	79.2	83.1	80.8	77.3	67.9	51.6	74.8	77.2	77.4	79	72.1	63.3	51
	20p	91.7	95.2	97.9	86.8	88	90.9	91.5	91.1	85	83.2	84.7	87.1	89.9	89.9	89.4	84.9
	30p	92.6	96.8	99.9	89.7	90.7	92.5	96.1	97.1	96.5	85.5	87.2	90.2	91.5	93.6	93.8	93.2
	40p	94.6	96.6	100	89.9	92.2	94.3	95.6	98.4	99.2	88.2	90.6	91.8	92.7	95.1	95.2	96.3
	60p	95.5	97.9	100	92.5	94.5	94.4	97.7	98.3	100	90.3	91.7	92.8	94.4	95	95.8	97.2
	100p	97.2	98.6	100	94.2	96.9	96.6	98.1	99	100	93.4	94.5	95.7	95.2	96.4	97.7	98.4

Figure 1. Percentage of correct model selection for BIC and AIC; four total predictors

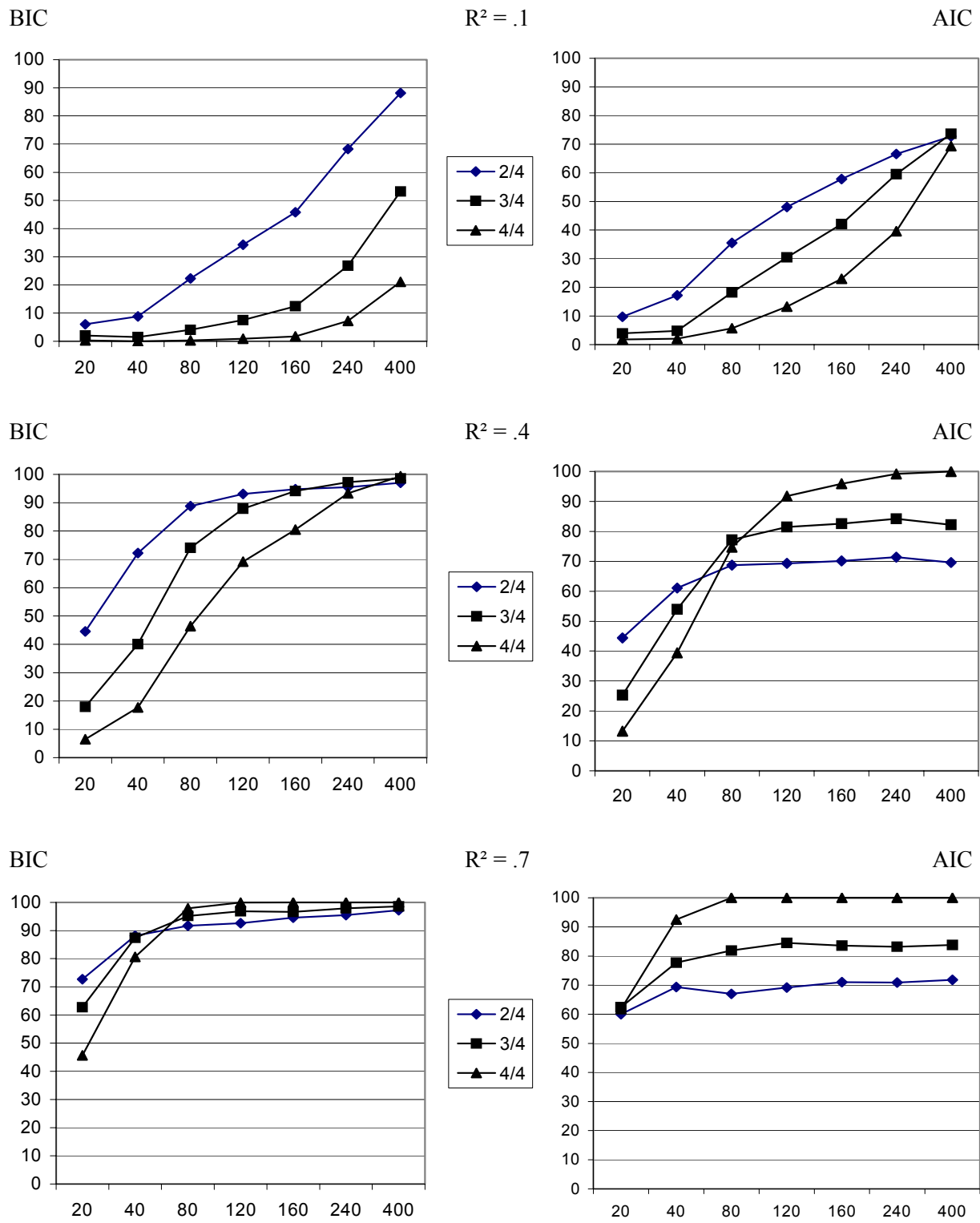


Figure 2. Percentage of correct model selection for BIC and AIC; seven total predictors

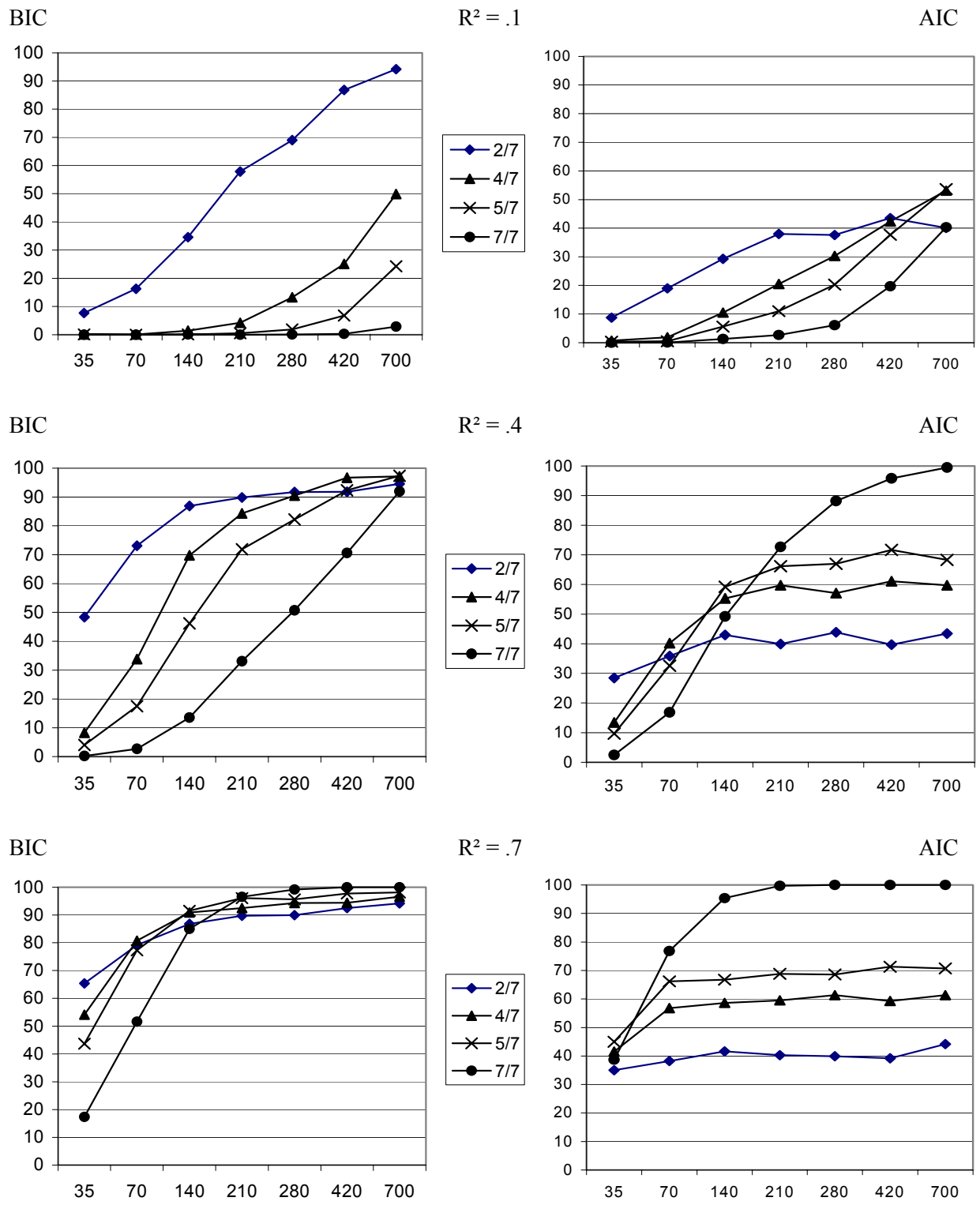
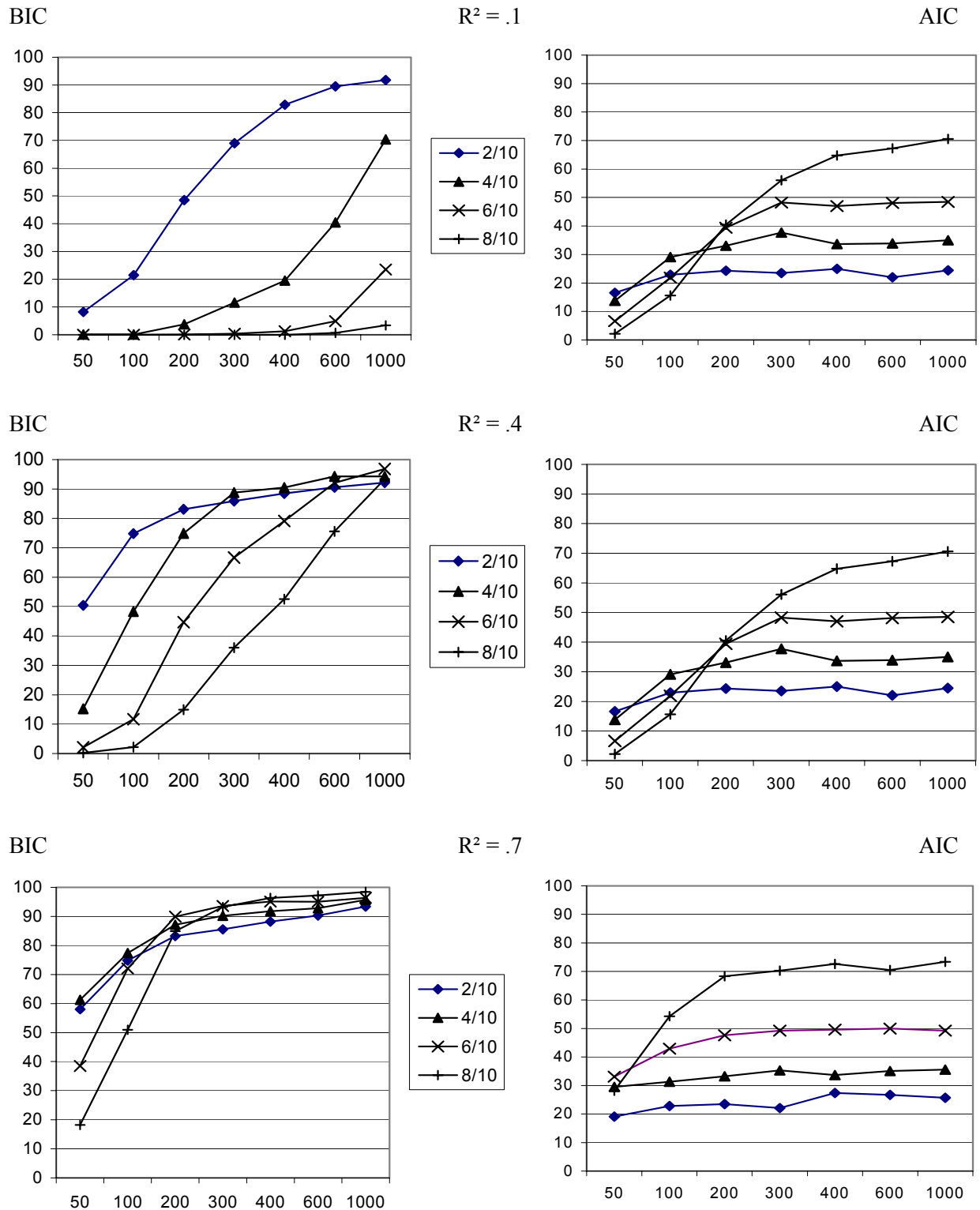


Figure 3. Percentage of correct model selection for BIC and AIC; ten total predictors





*JMASM Algorithms and Code*  
**JMASM4: Critical Values For Four Nonparametric And/Or Distribution-Free  
Tests Of Location For Two Independent Samples**

Bruce R. Fay  
Assessment & Evaluation  
Wayne County Regional Educational Service Agency

---

Researchers engaged in computer-intensive studies may need exact critical values, especially for sample sizes and alpha levels not normally found in published tables, as well as the ability to control ‘best-fit’ criteria. They may also benefit from the ability to directly generate these values rather than having to create lookup tables. Fortran 90 programs generate ‘best-conservative’ (bc) and ‘best-fit’ (bf) critical values with associated probabilities for the Kolmogorov-Smirnov test of general differences (bc), Rosenbaum’s test of location (bc), Tukey’s quick test (bc and bf) and the Wilcoxon rank-sum test (bc).

Key words: Kolmogorov-Smirnov test, Rosenbaum test, Tukey quick test; Wilcoxon rank-sum test.

---

### Introduction

Researchers, especially those engaged in Monte Carlo studies, may have a need for exact critical values over a wider range of sample sizes and/or alpha levels than are generally available from published tables. They may also benefit from the ability to generate the values directly, as opposed to creating lookup tables, and to control best-fit criteria. Fortran 90 programs that generate critical values for four nonparametric/distribution-free tests of location for two independent samples are presented. Included are the Kolmogorov-Smirnov test of general differences, Rosenbaum’s test of location, Tukey’s quick test and the Wilcoxon rank-sum test. The programs for Tukey’s test also generate ‘best-fit’ critical values and associated probabilities. The best-fit method could be adapted to the other programs.

### Tukey Quick Test

Tukey (1959) described a method for generating critical values for his *Two-Sample Test to Duckworth’s Specifications*, now commonly known as Tukey’s Quick Test. The test is both quick and compact, which makes it portable. The “rule of thumb” critical values, however, are not consistently ‘best-conservative’ or ‘best-fit’ to specific criteria.

### Test Description

Tukey’s (1959) test is quick in the sense that the method is easily remembered and the statistic, based on the combined length of extreme runs, easily calculated. The two samples are combined and ordered. For a two-sided test, if the overall maximum and minimum come from different groups, the statistic is the number of observations from the group with the global maximum that are greater than the greatest observation from the group with the global minimum plus the number of observations from the group with the global minimum that are less than the least observation from the group with the global maximum. If the global maximum and minimum are from the same group the statistic is generally taken to be zero. Tukey (1959) suggested dealing with ties (consequential,

---

Bruce R. Fay is an Assessment Consultant. He works with K-12 public schools in school accountability, accreditation, and assessment of student learning. Contact him at 30580 Springland St., Farmington Hills, MI 48334 or by e-mail at [bfay@twmi.rr.com](mailto:bfay@twmi.rr.com).

between-group) by counting each tied observation as  $\frac{1}{2}$  rather than 1. The one-sided (directional) test statistic is calculated just like the two-sided statistic with the additional requirement that the overall maximum observation is from the group that is expected to have the higher median under the alternative hypothesis (assuming a pure shift model). If not, the statistic is taken to be zero.

The test is compact in the sense that the critical values do not vary much with sample size, especially if the sample sizes are not too different. As such, they can also be easily committed to memory. For two-sided tests at nominal alpha levels of .10, .05, .02 and .01 (or one-sided tests at .05, .025, .01 and .005) the best-conservative critical values are 6, 7, 9 and 10 respectively with equal sample sizes from 9 to 24 per group. Tukey (1959) suggested that these critical values be used for all sample sizes as long as they were not too different. He noted, however, that under these conditions the test was not strictly conservative in the classical sense. He also gave relatively simple corrections to apply when the sample sizes were different, although not by too much. These corrections, however, still do not guarantee that the test will be strictly conservative, and add a level of complexity to the test that reduces both its quickness and compactness.

The best-fitting critical values for nominal alpha levels (1-sided) of .05, .025, .01, .005 (with a +10% tolerance) are 6, 7, 8 and 9 for equal samples sizes from 5 to 9 and 6, 7, 9, 10 for equal sample sizes from 11 to 30. Using 6, 7, 9, 10 as the critical values for all equal sample sizes is conservative for samples sizes less than 11 at .02 and .01 alpha levels (2-sided) but may be liberal up to +10% for other sample sizes and nominal alphas.

Quickness and compactness combine to make Tukey's (1959) test portable in the sense that everything needed to apply the test can be carried around in one's memory and the calculations can be performed mentally, or with pencil and paper. This simplicity is gained at the expense of some statistical power, but the practical power may be high. Tukey (1959) referenced a definition of practical power from Churchill Eisenhart (without formal citation) as "the product of the mathematical power by the probability that the procedure will be used" and noted that the practical power of a test might prove to be quite

high, in spite of lower statistical power, if it became widely used.

Because of its portability and potentially high practical power, Tukey (1959) referred to this test as a "pocket test" and proposed that it filled a particular niche, i.e., "as a footrule", "on the floor", or "in the field" to "indicate the weight of the evidence roughly." He recommended that more sensitive tests be used "if a delicate and critical decision is to be made."

#### Methodology for Generating Critical Values and Associated Probabilities

Tukey (1959) described in detail a method for generating strictly conservative, exact critical values. That method is implemented in the program modules presented here, along with a variation that produces best-fitting critical values to a specified tolerance level above nominal alpha.

Tukey's (1959) method involves building a table,  $A$ , that contains "a certain summation of binomial coefficients." Differences of pairs of entries from  $A$ , based on the sample sizes  $j$  and  $k$  and a parameter  $h$ , are compared to  ${}_nC_j$ , the number of combinations of  $n$  things taken  $j$  at a time, where  $n = j + k$ ,  $j \geq 1$ ,  $k \geq 1$ , and  $j \leq k$ . The differences  $A(k - h, j) - A(k, j - h)$  are formed starting with  $h = 1$  and counting up until the difference is less than  $(\text{nominal alpha}) \times ({}_nC_j)$ . The first such value of  $h$ , if one exists, is the best-conservative critical value for that pair of sample sizes and nominal alpha level. Additional details of the method are given in the comments that accompany the programs. Based on the use of integer\*8 and real\*8 variables, critical values and associated probabilities are generated for all combinations of sample sizes from (1, 1) to (30, 30) in increments of 1 for each sample. Tukey (1959) also presented asymptotic methods that may be appropriate for larger sample sizes.

The module that generates the critical values and associated probabilities contains two versions of the method and a subroutine for calculating combinations. The first version of the method generates strictly conservative critical values for one-sided tests at .05, .025, .01 and .005 nominal alpha levels. The second version generates 'best-fit' critical values for one-sided tests at the same nominal alpha levels. The 'best-fit' version allows critical values greater than nominal alpha so long as they do not exceed

nominal alpha by more than 10% and are closer to nominal alpha than the nearest value that is less than nominal alpha. The +10% tolerance is based on a definition of robustness due to Bradley (1978).

#### Rosenbaum's Test of Location

Rosenbaum (1953, 1954) described tests for dispersion and location based on Wilks (1942) and gave tables of critical values. Rosenbaum (1965) revisited these tests, comparing them to other tests that had arisen in the intervening decade. Neave & Worthington (1988) described the location form of the test as particularly well suited to situations in which spread is expected to increase with an increase in the median and gave a method for generating critical values. Their method is the basis for the programs presented here. Rosenbaum's (1954) test is quick and relatively compact, which makes it somewhat portable.

#### Test Description

The test is quick in the sense that the method is easily remembered and the statistic, based on the length of an extreme run, easily calculated. The two samples are combined and ordered. For a two-sided test, the statistic is taken as the number of observations from the group with the overall maximum that exceeds the maximum value of the other group. One way to deal with consequential (between-group) ties is to count each observation as  $\frac{1}{2}$  rather than 1. Another method is to average the values of the statistic arrived at by resolving the ties in all possible ways. The later technique, however, causes the test to lose some of its portability, at least for larger sample sizes. The one-sided (directional) test statistic is calculated just like the two-sided statistic with the additional requirement that the overall maximum observation is from the group that is expected to have the higher median under the alternative hypothesis (assuming a pure shift model). If not, the statistic is taken to be zero.

The test is compact in the sense that the critical values do not vary much with sample size, especially if the sample sizes are not too different. As such, they can also be easily committed to memory. For two-sided tests at nominal alpha levels of .10, .05, .02 and .01 (or one-sided tests at .05, .025, .01 and .005) the best-conservative

critical values are 5, 6, 7 and 8 respectively for equal sample sizes from 27 to 50 per group. Critical values of 5, 6, 7, and 8 can be used for equal sample sizes from 20 to 50, and critical values of 4, 5, 6 and 7 for equal sample sizes from 5 to 19, if one is willing to accept results that are not strictly conservative in all cases, and somewhat overly conservative in others. Under these conditions the test can be considered compact. Quickness and compactness combine to make the test portable as previously described.

#### Methodology for Generating Critical Values and Associated Probabilities

Neave & Worthington (1988) described a method for generating strictly conservative, exact critical values. Their method is implemented in the program modules presented here to calculate the critical values for one-sided tests at .05, .025, .01 and .005 nominal alpha levels.

Neave & Worthington (1988) calculated the probability of a run of  $h$  values from a sample of size  $m$  out of a combined sample of size  $N = m + n$ , where  $n$  is the size of the other group, using the formula:

$$\frac{m!(N-h)!}{N!(m-h)!} = \frac{m}{N} \times \frac{m-1}{N-1} \times \dots \times \frac{m-h+1}{N-m+1}. \quad (1)$$

The value of  $h$  associated with the largest such probability that is less than or equal to nominal alpha is the critical value for a given  $m$  and  $n$ . Thus all critical values are best-conservative with  $\text{pr}(\text{CV}) \leq \text{nominal alpha}$ . Additional details of the method are given in the comments that accompany the programs. Based on the use of integer\*8 and real\*8 variables, critical values and associated probabilities are generated for all combinations of sample sizes from (1, 1) to (50, 50) in increments of 1 for each sample.

#### Kolmogorov-Smirnov Test of General Differences

Kim and Jennrich (1970, 1973) cited Smirnov (1939) as introducing the criterion  $D_{mn}$  for the two-sample problem. As the name implies, the test is sensitive to general differences between two populations and is often used as a 2-sided test. Neave and Worthington (1988) pointed out, however, that the test functions quite well as a directional (1-sided) test, especially against a pure

shift alternative. Kim and Jennrich (1970, 1973) provided a brief review of work on approximate and exact distributions of the statistic and resultant critical values under the null hypothesis leading up to their method and tables.

Test Description

The 2-sided test is conducted by constructing and then comparing the empirical cumulative distributions,  $S_m(x)$  and  $S_n(x)$ , of two samples of size  $m$  and  $n$  ( $m \leq n$  without loss of generality) and then computing the criterion as  $D_{mn} = \sup | S_m(x) - S_n(x) |$  over all  $x$ . The null hypothesis is that the two samples are drawn from identical (continuous) populations  $F_m(x)$  and  $F_n(x)$  (of any shape). The alternative hypothesis is that the samples were drawn from two populations that differ in some way. For a 1-sided test under a pure shift model, the criterion is taken to be  $D_{mn}^+$  or  $D_{mn}^-$ , where  $D_{mn}^+ = \max [ S_m(x) - S_n(x) ] \geq 0$  and  $D_{mn}^- = \min [ S_m(x) - S_n(x) ] \leq 0$ . The choice depends on which sample is presumed to come from the population with the higher median under the alternative hypothesis. If the alternative hypothesis is that the samples came from populations with cumulative distributions such that  $F_n(x) \geq F_m(x)$  then  $S_n(x)$  will lie to the right of  $S_m(x)$ . Thus,  $S_m(x)$  will rise faster than  $S_n(x)$  and lie above it for any given value of  $x$ . This makes  $D_{mn}^+$  the correct choice of criterion in this case.

Methodology for Generating Critical Values and Associated Probabilities

The Kim and Jennrich (1970, 1973) method of generating critical values for the Kolmogorov- Smirnov test is based on the work of Kim (1969) which, in turn, was an extension of the successive recursion relation of Massey (1951). Their method calculates:

$$P\left(D_{mn} \leq \frac{c}{mn}\right) = U(m, n) \tag{2}$$

where

$$U(i, j) = \frac{i}{i+n} C(i, j) [U(i, j-1) + U(i-1, j)] \tag{3}$$

and

$$C(i, j) = \begin{cases} 1 & \text{if } \left| \frac{i}{m} - \frac{j}{n} \right| \leq \frac{c}{mn} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

subject to initial condition

$$U(i, j) = \binom{i+n}{i}^{-1} C(i, j), \text{ when } i \bullet j = 0. \tag{5}$$

Kim and Jennrich (1970, 1973) provided a FORTRAN IV function subroutine *ASKCDF*( $M, N, D, U$ ) that returned the probability of  $D (= c/mn)$  for sample sizes  $m$  and  $n$  by calculating  $U(m, n)$  as above. The  $U$  referenced in their function subroutine argument list, however, was merely a working storage vector of at least length  $N+1$ . In the Fortran 90 implementation of *ASKCDF* that follows, the working storage vector argument has been eliminated and replaced in the code with an allocatable array. A subroutine calculates  $D = c/mn$  for  $c = (1, mn, 1)$  for each combination of  $n = (1, 50, 1)$  and  $m = (1, n, 1)$  and calls *ASKCDF* for each value of  $D$  to obtain the probability and tests it against various nominal alpha levels.

Wilcoxon Rank-sum Test

Wilcoxon (1945) introduced the non-parametric/distribution-free test based on a sum of ranks that bears his name. Wilcoxon (1946, 1947) expanded on this work, followed by Mann and Whitney (1947), who described a test that turned out to be equivalent to the rank-sum test. The Wilcoxon-Mann-Whitney test is probably the best known of the nonparametric/distribution-free procedures. However, the early work of both Wilcoxon and Mann-Whitney provided only limited critical values. Additional work on both exact and approximate critical values and significance probabilities followed these seminal articles, e.g. Fix & Hodges (1955).

Jacobson (1963) provided a nice synopsis of critical value tables and work-to-date with an extensive bibliography. Wilcoxon and Wilcox (1964, revised 1968) provided a workable method for generating critical values and probability levels. This work subsequently appeared in Wilcoxon, Katti and Wilcox (1970, revised 1973)

and forms the basis for the programs presented here.

#### Test Description

The Wilcoxon rank-sum version of the test is conducted by combining the observations from two samples. The combined samples are then ranked while keeping track of the original group membership. The ranks from one of the groups are then summed to form the statistic. Which group to sum for a 1-sided test depends on the critical value tables that are available (lower-tail, upper tail, or both) and on which group is expected to have the least (or greatest) ranks under the alternative hypothesis. For example, if lower tail critical values are available, and the alternative hypothesis is that sample *B* comes from a population that is greater than the population from which sample *A* was obtained, then sample *A* will tend to have the lower ranks, and the sum of those ranks would be taken as the statistic. For a two-sided test, one would form the sum of the ranks of both samples and test the resulting values against the critical value, taking the test to be significant if either comparison so indicated.

#### Methodology for Generating Critical Values and Associated Probabilities

Although critical values are readily available for the Wilcoxon rank-sum test and Mann-Whitney *U* test, the probability levels are not as accessible. The method of Wilcoxon, Katti and Wilcox (1970, 1973) proceeds along the following lines given samples *M* and *N* from two continuous populations,  $F_m(x)$  and  $G_n(x)$  of size *m* and *n* respectively,  $m \leq n$  without loss of generality. The minimum sum of ranks for sample *M* is  $m(m+1)/2$ . Thus the sum of ranks in general for sample *M* is:

$$\frac{m(m+1)}{2} + U \text{ where } U \in I, U \geq 0. \quad (6)$$

The number of ways,  $f(U)$ , of obtaining a specific rank sum *U*, is the coefficient of  $t^U$  in the expansion of the generating function, in powers of *t*, given by:

$$g(t) = \prod_{i=1}^n \frac{(1-t^{m+i})}{(1-t^i)}. \quad (7)$$

The total number of ways of obtaining any rank sum in this situation is:

$$T = \binom{m+n}{n}. \quad (8)$$

Given  $F_m(x) \equiv G_n(x)$ , the probability of obtaining *U* is given by:

$$pr(U) = \frac{f(U)}{T}. \quad (9)$$

In turn,  $f(U)$  can be found from:

$$f(U) = \frac{1}{U} \sum_{i=0}^{U-1} f(i) z_{U-i-1} \quad (10)$$

for ( $U = 1, 2, 3, \dots$ ) and with  $f(0) = 1$

In order to evaluate equation (10) it is necessary to find the values of *z*. Subroutine CV\_WRSJ4\_init in module CVWRSJmod includes the code for generating the values of *z*.

#### Source Code and Computing Platforms

All source code provided here is Fortran 90 free format. For each of the four tests there is a module that contains the critical value generation subroutines and functions and a main program that can be used with that module to generate printed tables of critical values and probabilities. The programs were developed on a 500 MHz AMD Athlon-based system using Compaq Visual Fortran 6.6 and tested on systems with Intel Pentium III and Pentium IV Xeon processors. The programs execute reasonably quickly on all of these systems. Even with integer\*8 and real\*8 variables these programs can run into arithmetic overflow problems, thus limiting the range of sample sizes for which critical values and probabilities can be generated.

## References

- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Fix, E. and Hodges, J. L. Jr. (1955). Significance probabilities of the Wilcoxon test. *Annals of Mathematical Statistics*, 26, 301-312.
- Jacobson, J. E. (1963). The Wilcoxon two-sample statistic: Tables and bibliography. *Journal of the American Statistical Association*, 58, 1086-1103.
- Kim, P. J. (1969). On the exact and approximate sampling distribution of the two sample Kolmogorov-Smirnov criterion  $D_{mn}$ ,  $m \leq n$ . *Journal of the American Statistical Association*, 64: 1625-1637.
- Kim, P. J. & Jennrich, R. I. (1970, 1973). Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion,  $D_{mn}$ ,  $m \leq n$ . *Selected Tables in Mathematical Statistics, Volume I* (1970, 2<sup>nd</sup> printing with revisions, 1973), 77-170. Harter, H. L. & Owen, D. B., coeditors, Providence, RI: American Mathematical Society (edited by the Institute of Mathematical Statistics).
- Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50-60.
- Massey, F. J. Jr. (1951). The distribution of the maximum deviation between two sample cumulative step functions. *Annals of Mathematical Statistics*, 22, 125-128.
- Neave, H. R. & Worthington, P. L. B. (1988). *Distribution-free tests*. Unwin Hyman Ltd.
- Rosenbaum, S. (1953). Tables for a nonparametric test of dispersion. *Annals of Mathematical Statistics*, 24, 663-668.
- Rosenbaum, S. (1954). Tables for a nonparametric test of location. *Annals of Mathematical Statistics*, 25, 146-150.
- Rosenbaum, S. (1965). On some two-sample non-parametric tests. *Journal of American Statistical Association*, 60, 1118-1126.
- Smirnov, N. V. (1939). Estimating the deviation between the empirical distribution functions of two independent samples. *Bulletin de l'Universite' de Moscou*, 2(2,3).
- Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19, 279-281.
- Tukey, J. W. (1959). A quick, compact, two-sample test to Duckworth's specifications. *Technometrics*, 1(1), 31-48.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.
- Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *Journal of Economic Entomology*, 39(2), 269.
- Wilcoxon, F. (1947). Probability levels for individual comparisons by ranking methods. *Biometrics*, 3, 119-122.
- Wilcoxon, F. & Wilcox, R. A. (1964). Some rapid approximate statistical procedures. Pearl River, NY: Lederle Laboratories Division, American Cyanamid Company. (Originally prepared and distributed in cooperation with the Department of Statistics, The Florida State University, Tallahassee, FL. and revised, 1968).
- Wilcoxon, F., Katti, S. K., & Wilcox, R. A. (1970, 1973). Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected Tables in Mathematical Statistics, Volume I* (1970, 2<sup>nd</sup> printing with revisions, 1973), 171-259. Harter, H. L. & Owen, D. B., coeditors, Providence, RI: American Mathematical Society (edited by the Institute of Mathematical Statistics).
- Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Annals of Mathematical Statistics*, 13, 400-409.

## Programs

## Tukey's (1959) Two-sample Test to Duckworth's Specifications (Tukey's Quick Test)

## Main program for printing tables

```

! *****
! program:  CVTQTJ.exe
! source:   CVTQTJ.f90
! author:   Bruce R. Fay
! date:    17 Oct 2002 17:32 EDT
! purpose:  Test harness for critical value modules for Tukey's Quick
!           test of location to Duckworth's specifications
! desc:     Prints tables of critical values with associated probabilities.
! *****
program CVTQTJ
use CVTQTJmod
implicit none
! DECLARE LOCAL VARIABLES
integer :: i, j, LU1, LU2, ios, testnum
integer, dimension(:) :: CVi(4)
real*8, dimension(:) :: PVr(4)
! GET USER INPUTS
write(*,*) "Program CVTQTJ.exe by Bruce R. Fay"
write(*,*) "Critical values for Tukey's Quick Test"
write(*,*)
write(*,*) "Creates output files CVTQTJbc_.txt and CVTQTJbf_.txt"
write(*,*) "in current directory."
write(*,*)
write(*,*) "Select one of the following:"
write(*,*)
write(*,*) " 0 - to exit program"
write(*,*) " 1 - to generate CV/PV tables"
write(*,*)
Do
  read(*,*) testnum
  If ( (testnum >= 0).and.(testnum <= 1) ) EXIT
  write(*,*) "enter 0 to exit, 1 to run"
End Do
If (testnum == 0) GOTO 9999 ! check for user termination
! OPEN FILES FOR OUTPUT
LU1 = 8
open(unit=LU1, file='CVTQTJbc_.txt', iostat=ios)
IF (ios > 0 ) then
  write(*,*) "Error opening file 'CVTQTJbc_.txt' "
  GOTO 9999
End if
LU2 = 9
open(unit=LU2, file='CVTQTJbf_.txt', iostat=ios)
IF (ios > 0 ) then
  write(*,*) "Error opening file 'CVTQTJbf_.txt' "
  GOTO 9999
End if
! DEFINE OUTPUT FORMATS
100 format(" 1-tailed CVs at stated alpha levels")
200 format(" n1 n2 - .05 - -.025 - -.01 - -.005 - | &
           & - .05 - -.025 - -.01 - -.005 -")

```

```

300 format(2I3,4I8,3x,4F8.4)
! CREATE BEST-CONSERVATIVE TABLES
write(LU1,*) "Program CVTQTJ by Bruce R. Fay"
write(LU1,*)
write(LU1,*) "Tukey's quick test of location for two independent samples,"
write(LU1,*) "best-conservative critical values generated based on"
write(LU1,*) "Tukey (1959) using CVTQTJbc() in CVTQTJmod."
write(LU1,*)
call CV_TQTJbc_init      ! generate the BC CV/PV tables
write(LU1,100)          ! print header information
write(LU1,*)
write(LU1,200)          ! print column headers for this format
write(LU1,*)
Do i = 1,30              ! output the tables to file
  Do j = i,30
    call CV_TQTJbc(i,j,CVi,PVr)
    write(LU1,300) i,j,CVi(1:4),PVr(1:4)
  End Do
  write(LU1,*)
End Do
! CREATE BEST-FIT TABLES
write(LU2,*) "Program CVTQTJ by Bruce R. Fay"
write(LU2,*)
write(LU2,*) "Tukey's quick test of location for two independent samples."
write(LU2,*) "Best-fitting critical values generated based on Tukey (1959)"
write(LU2,*) "using CVTQTJbf() in CVTQTJmod, where best-fit is defined as"
write(LU2,*) "pr <= alpha + 10% when this probability is closer to alpha"
write(LU2,*) "than the first available CV with pr < alpha."
write(LU2,*)
call CV_TQTJbf_init     ! generate the BF CV/PV tables
write(LU2,100)          ! print header information
write(LU2,*)
write(LU2,200)          ! print column headers for this format
write(LU2,*)
Do i = 1,30              ! output the tables to file
  Do j = i,30
    call CV_TQTJbf(i,j,CVi,PVr)
    write(LU2,300) i,j,CVi(1:4),PVr(1:4)
  End Do
  write(LU2,*)
End Do
! CLOSE FILES
close(unit=LU1, status='keep', iostat=ios)
If (ios > 0) then
  write(*,*) "Error closing file 'CVTQTJbc_.txt' "
End If
close(unit=LU2, status='keep', iostat=ios)
If (ios > 0) then
  write(*,*) "Error closing file 'CVTQTJbf_.txt' "
End If
9999 stop
end program CVTQTJ

```



## Module for generating critical values and probabilities

```

! *****
! module:   CVTQJTJmod
! source:   CVTQJTJmod.f90
! based on: Tukey (1959) A quick, compact, two-sample test to Duckworth's
!           specifications, Technometrics Vol. 1 No. 1 (Feb) pgs.31-48,
!           method for generating exact critical values.
! author:   Bruce R. Fay
! date:     17 Oct 2002 19:03 EDT
! purpose:  Provide the exact critical values for Tukey's Quick Test for
!           2-independent-samples, both best-conservative and best-fit.
! desc:     Generates the CVTs and PVTs on initialization and provides
!           an entry point that returns up to four critical values based
!           on the incoming values of n1 and n2. Checks are made that
!           n1, n2 are in the appropriate range and relationship for the
!           tables with 1 <= n1 <= n2 <= 30.
! Notes:    Best-conservative values are those for which pr(h) <= nominal
!           alpha. Best-fit CVs are generated by the same method but with
!           pr(h) <= alpha+10% if pr(h+1) < alpha and is further from alpha
!           than pr(h).
! *****
module CVTQJTJmod
implicit none
private
public :: CV_TQTJbc_init, CV_TQTJbc, CV_TQTJbf_init, CV_TQTJbf, N_c_m
contains
! *****
subroutine CV_TQTJbc_init
! INTERFACE
! There are no arguments for CV_TQTJbc_init. The calling routine must call
! this subroutine once to build the CV and PV tables prior to calling
! CV_TQTJbc() to obtain critical values for specific n1, n2. Calling routine
! must also declare an integer vector of length 4 and a real*8 vector of
! length 4 and pass them into receive the critical values and their
! associated probability values. For entry CV_TQTJbc(s1,s2,CV,PV):
!   s1   :: sample size for 1st group ( <= s2 )
!   s2   :: sample size for 2nd group
!   CV   :: critical values vector (length 4)
!   PV   :: probability values vector (length 4)
! DECLARE DUMMY VARIABLES
integer, intent(in) :: s1, s2
integer, intent(out), dimension(:) :: CV
real*8, intent(out), dimension(:) :: PV
! DESCRIPTION
! At entry CV_TQTJbc(), for s1 <= s2, returns up to four critical values,
! if available, in vector CV(:), as follows:
!   CV(1) = 1-tailed alpha .05 (2-tailed alpha .10)
!   CV(2) = 1-tailed alpha .025 (2-tailed alpha .05)
!   CV(3) = 1-tailed alpha .01 (2-tailed alpha .02)
!   CV(4) = 1-tailed alpha .005 (2-tailed alpha .01)
! The actual 1-tailed probabilities corresponding to the above CVs are
! returned in PV(1:4). If a critical value is not available, a -1 is
! returned instead, with associated probability zero. Critical values may
! not be available because s1 and s2 are a) too small, b) too large, or

```

```

! c) too different. Unequal s1, s2 are supported for 1 <= s1 <= s2 <= 30.
! DECLARE LOCAL VARIABLES
integer :: h, n1, n2, v1, v2, w1, w2
integer (kind=8) :: wv1, wv2
integer (kind=8), dimension(30,30), save :: CVTbc05, CVTbc025
integer (kind=8), dimension(30,30), save :: CVTbc01, CVTbc005
integer (kind=8), dimension(0:30,1:30), save :: Atbl
integer (kind=8) :: comb, A1, A2, Adiff
integer (kind=8), parameter :: zero=0, one=1, two=2
real (kind=8), dimension(30,30), save :: PVTbc05, PVTbc025, PVTbc01, PVTbc005
real (kind=8), parameter :: m05=0.050, m025=0.025, m01=0.01, m005=0.005
real (kind=8) :: c05, c025, c01, c005, rcomb, rdifff
logical :: fnd05, fnd025, fnd01, fnd005
! Build the A table
!
!           Column(w)
!           -2    -1     0     1     2     3     4     ...    30
!           -----
!   -1|     0     0 |     0 |     0     0     0     0     ...     0
!     0|     0     0 |     0 |     0     0     0     0     ...     0
!           |-----+-----+-----
!     1|     0     1 |     1 |     1     1     1     1     ...     1
!           |-----+-----+-----
! Row 2|     1     1 |     2 |     3     4     5     6     ...    32
! (v) 3|     1     2 |     4 |     7    11    16    22     ...     .
!     4|     2     4 |     8 |    15    26    42    64     ...     .
!     5|     4     8 |    16 |    31    57    99   163     ...     .
!     6|     8    16 |    32 |     .     .     .     .     ...     .
!     .|    16    32 |     . |     .     .     .     .     ...     .
!     .|    32     . |     . |     .     .     .     .     ...     .
!     .|     .     . |     . |     .     .     .     .     ...     .
!    30|     .     . |     . |     .     .     .     .     ...     .
!           536870912
!
! Note: The A table is only built for columns 0 to 30 and rows 1 to 30. All
! entries for rows less than one are zero and all entries for columns
! less than zero (with rows of 1 or more) can be determined by direct
! formula (see code).
Atbl(0:30,1) = one      ! first row, all columns, entries = 1
Do v1 = 2,30           ! first (zero) column, row entries are 2^(row-1)
  Atbl(0,v1) = two**(v1-1)
End Do
Do v1 = 2,30           ! previous column same row + same column previous row
  Do w1 = 1,30
    Atbl(w1,v1) = Atbl(w1-1,v1) + Atbl(w1,v1-1)
  End Do
End Do
CVTbc05 = -1      ! initialize the CV tables to -1 (indicates no valid entry)
CVTbc025 = -1
CVTbc01 = -1
CVTbc005 = -1
PVTbc05 = 0.0    ! initialize the PV tables to 0.0 (indicates no valid entry)
PVTbc025 = 0.0
PVTbc01 = 0.0
PVTbc005 = 0.0
! Determine the critical values and associated actual probabilities
Do n1 = 1,30        ! n1 for CV/PV tables
  Do n2 = n1,30     ! n2 for CV/PV tables

```

```

fnd05 = .false.      ! reset found flags for each alpha level
fnd025 = .false.
fnd01 = .false.
fnd005 = .false.
comb = N_c_m(n1,n2) ! get the number of combinations for n1 and n2
rcomb = real(comb)
c05 = rcomb * m05   ! calculate the comparison values for each alpha
c025 = rcomb * m025
c01 = rcomb * m01
c005 = rcomb * m005
Do h = 1,(n1+n2)    ! h will be the CV if/when we find the right one
  w1 = n2-h        ! Find A1 as Atbl(n2-h,n1)
  v1 = n1          ! since n1 >= 1, v is a valid row for Atbl
  wv1 = w1 + v1   ! = n1 + n2 - h
  If (w1 >= 0) then ! it's OK to use the Atbl to get A1
    A1 = Atbl(w1,v1)
  Else             ! calculate A1 by formula
    If (wv1 > 0) then ! w < 0, v > 0, |v| > |w|
      A1 = two**(wv1-1)
    Else If (wv1 == 0) then ! w = -v
      A1 = one
    Else If (wv1 < 0) then ! w < 0, v > 0, |v| < |w|
      A1 = zero
    End If
  End If
  v2 = n1-h        ! Find A2 as Atbl(n2,n1-h)
  w2 = n2          ! since n2 >= 1, w is a valid column for Atbl
  If(v2 >= 1) then ! valid row for Atbl
    A2 = Atbl(w2,v2)
  Else
    A2 = zero
  End If
  Adiff = A1 - A2
  rdifff = real(Adiff)
  If ( (rdifff <= c05).and.(.not.fnd05) ) then
    CVTbc05(n1,n2) = h
    PVTbc05(n1,n2) = rdifff/rcomb
    fnd05 = .true.
  End If
  If ( (rdifff <= c025).and.(.not.fnd025) ) then
    CVTbc025(n1,n2) = h
    PVTbc025(n1,n2) = rdifff/rcomb
    fnd025 = .true.
  End If
  If ( (rdifff <= c01).and.(.not.fnd01) ) then
    CVTbc01(n1,n2) = h
    PVTbc01(n1,n2) = rdifff/rcomb
    fnd01 = .true.
  End If
  If ( (rdifff <= c005).and.(.not.fnd005) ) then
    CVTbc005(n1,n2) = h
    PVTbc005(n1,n2) = rdifff/rcomb
    fnd005 = .true.
  End If
  If (fnd05.and.fnd025.and.fnd01.and.fnd005) exit
End Do
End Do

```

```

End Do
Return
! -----
entry CV_TQTJbc(s1,s2,CV,PV)
CV(:) = -1      ! initialize all return CVs to 'not available'
PV(:) = 0.0    ! initialize all return PVs to 'not available'
If ((1<=s1).and.(s1<=30).and.(1<=s2).and.(s2<=30).and.(s1<=s2)) then
  CV(1) = CVTbc05(s1,s2)
  CV(2) = CVTbc025(s1,s2)
  CV(3) = CVTbc01(s1,s2)
  CV(4) = CVTbc005(s1,s2)
  PV(1) = PVTbc05(s1,s2)
  PV(2) = PVTbc025(s1,s2)
  PV(3) = PVTbc01(s1,s2)
  PV(4) = PVTbc005(s1,s2)
End If
Return
! -----
end subroutine CV_TQTJbc_init
! *****
subroutine CV_TQTJbf_init
! see subroutine CV_TQTJbc_init above for documentation and comments
! DECLARE DUMMY VARIABLES
integer, intent(in) :: s1, s2
integer, intent(out), dimension(:) :: CV
real*8, intent(out), dimension(:) :: PV
! DECLARE LOCAL VARIABLES
integer :: h, n1, n2, v1, v2, w1, w2
integer (kind=8) :: CV1tmp, CV2tmp, CV3tmp, CV4tmp, wv1, wv2
integer (kind=8), dimension(30,30), save :: CVTbf05, CVTbf025
integer (kind=8), dimension(30,30), save :: CVTbf01, CVTbf005
integer (kind=8), dimension(0:30,1:30), save :: Atbl
integer (kind=8) :: comb, A1, A2, Adiff
integer (kind=8), parameter :: two=2
real (kind=8), dimension(30,30), save :: PVTbf05, PVTbf025, PVTbf01, PVTbf005
real (kind=8), parameter :: m05=0.05, m025=0.025, m01=0.01, m005=0.005
real (kind=8), parameter :: m055=0.055, m0275=0.0275
real (kind=8), parameter :: m011=0.011, m0055=0.0055
real (kind=8) :: c05, c025, c01, c005, c055, c0275, c011, c0055, rcomb, rdifff
real (kind=8) :: ptmp, PV1tmp, PV2tmp, PV3tmp, PV4tmp
logical :: fnd05, fnd025, fnd01, fnd005
! BUILD THE A TABLE
Atbl(0:30,1) = 1 ! first row
Do v1 = 1,30 ! first column
  Atbl(0,v1) = two**(v1-1)
End Do
Do v1 = 2,30 ! previous column same row + same column previous row
  Do w1 = 1,30
    Atbl(w1,v1) = Atbl(w1-1,v1) + Atbl(w1,v1-1)
  End Do
End Do
CVTbf05 = -1 ! initialize the CV tables to -1 (indicates no valid entry)
CVTbf025 = -1
CVTbf01 = -1
CVTbf005 = -1
PVTbf05 = 0.0 ! initialize the PV tables to 0.0 (indicates no valid entry)
PVTbf025 = 0.0

```

```

PVTbf01 = 0.0
PVTbf005 = 0.0
! Determine the critical values and associated actual probabilities
Do n1 = 1,30
  Do n2 = n1,30
    fnd05 = .false. ! reset found flags for each alpha level
    fnd025 = .false.
    fnd01 = .false.
    fnd005 = .false.
    comb = N_c_m(n1,n2) ! get the number of combinations for n1 and n2
    rcomb = real(comb)
    c05 = rcomb * m05 ! calculate the comparison values for each alpha
    c025 = rcomb * m025
    c01 = rcomb * m01
    c005 = rcomb * m005
    c055 = rcomb * m055 ! comparison values for alpha + 10%
    c0275 = rcomb * m0275
    c011 = rcomb * m011
    c0055 = rcomb * m0055
    PV1tmp = 1.0 ! initialize temporary probability values
    PV2tmp = 1.0
    PV3tmp = 1.0
    PV4tmp = 1.0
    Do h = 1,(n1+n2)
      w1 = n2-h
      v1 = n1
      wv1 = w1 + v1
      If (w1 >= 0) then
        A1 = Atbl(w1,v1)
      Else
        If (wv1 > 0) then
          A1 = 2**(wv1-1)
        Else If (wv1 == 0) then
          A1 = 1
        Else If (wv1 < 0) then
          A1 = 0
        End If
      End If
      w2 = n2
      v2 = n1-h
      If (v2 >= 1) then
        A2 = Atbl(w2,v2)
      Else
        A2 = 0
      End If
      Adiff = A1 - A2
      rdiff = real(Adiff)
      If((c05 < rdiff).and.(rdiff <= c055).and.(.not.fnd05)) then
        CV1tmp = h
        PV1tmp = rdiff/rcomb
      Else If((rdiff <= c05).and.(.not.fnd05)) then
        ptmp = rdiff/rcomb
        If((.05 - ptmp) <= (PV1tmp - .05)) then
          CVTbf05(n1,n2) = h
          PVTbf05(n1,n2) = ptmp
        Else
          CVTbf05(n1,n2) = CV1tmp
        End If
      End If
    End Do
  End Do
End Do

```

```

    PVTbf05(n1,n2) = PV1tmp
  End If
  fnd05 = .true.
End If
If((c025 < rdiff).and.(rdiff <= c0275).and.(.not.fnd025)) then
  CV2tmp = h
  PV2tmp = rdiff/rcomb
Else If((rdiff <= c025).and.(.not.fnd025)) then
  ptmp = rdiff/rcomb
  If((.025 - ptmp) <= (PV2tmp - .025)) then
    CVTbf025(n1,n2) = h
    PVTbf025(n1,n2) = ptmp
  Else
    CVTbf025(n1,n2) = CV2tmp
    PVTbf025(n1,n2) = PV2tmp
  End If
  fnd025 = .true.
End If
If((c01 < rdiff).and.(rdiff <= c011).and.(.not.fnd01)) then
  CV3tmp = h
  PV3tmp = rdiff/rcomb
Else If((rdiff <= c01).and.(.not.fnd01)) then
  ptmp = rdiff/rcomb
  If((.01 - ptmp) <= (PV3tmp - .01)) then
    CVTbf01(n1,n2) = h
    PVTbf01(n1,n2) = ptmp
  Else
    CVTbf01(n1,n2) = CV3tmp
    PVTbf01(n1,n2) = PV3tmp
  End If
  fnd01 = .true.
End If
If((c005 < rdiff).and.(rdiff <= c0055).and.(.not.fnd005)) then
  CV4tmp = h
  PV4tmp = rdiff/rcomb
Else If((rdiff <= c005).and.(.not.fnd005)) then
  ptmp = rdiff/rcomb
  If((.005 - ptmp) <= (PV4tmp - .005)) then
    CVTbf005(n1,n2) = h
    PVTbf005(n1,n2) = ptmp
  Else
    CVTbf005(n1,n2) = CV4tmp
    PVTbf005(n1,n2) = PV4tmp
  End If
  fnd005 = .true.
End If
If (fnd05.and.fnd025.and.fnd01.and.fnd005) exit
End Do
End Do
Return
! -----
entry CV_TQTJbf(s1,s2,CV,PV)
CV(:) = -1      ! initialize all return CVs to 'not available'
PV(:) = 0.0     ! initialize all return PVs to 'not available'
If ((1<=s1).and.(s1<=30).and.(1<=s2).and.(s2<=30).and.(s1<=s2)) then
  CV(1) = CVTbf05(s1,s2)

```

```

CV(2) = CVTbf025(s1,s2)
CV(3) = CVTbf01(s1,s2)
CV(4) = CVTbf005(s1,s2)
PV(1) = PVTbf05(s1,s2)
PV(2) = PVTbf025(s1,s2)
PV(3) = PVTbf01(s1,s2)
PV(4) = PVTbf005(s1,s2)
End If
Return
! -----
end subroutine CV_TQTJbf_init
!*****
function N_c_m(a,b) result(F)
! Calculates number of combinations, 'N chose m' or nCm where
! N = a+b and m = a (equivalent to m = b). The formula is
!  $N!/(m!(N-m)!) = (a+b)!/(a!b!) =$ 
!  $[1*2*...*b*(b+1)*...*(a+b)]/[(1*2*...*a)*(1*2*...*b)]$ 
! This is equivalent to  $[(b+1)(b+2)...(b+a)]/[a!]$  or
!  $[(b+1)(b+2)...(b+a)]/[1*2*...*a]$ , which is implemented here.
! This computation is particularly efficient if  $a \leq b$ , as it is in
! subroutines CV_TQTJbc_init and CV_TQTJbf_init above. Both a and b must
! be  $\geq$  zero, otherwise the function returns with value -1 to indicate an
! error.
! DECLARE DUMMY VARIABLES
integer, intent(in) :: a, b
! DECLARE LOCAL VARIABLES
integer :: i
integer (kind=8) :: C, F, num
! VARIABLE DEFINITIONS
! a      :: number of items in first group
! b      :: number of items in second group
! C      :: accumulator for number of combinations
! F      :: function result
! i      :: loop variable
! num    :: numerator factor for combinations computation
If((a $\geq$ 0).and.(b $\geq$ 0)) then ! both inputs non-negative
  If((a $\geq$ 1).and.(b $\geq$ 1)) then ! both inputs  $> 0$ , proceed
    C = 1
    Do i = 1,a
      num = i + b
      C = (C * num) / i
    End Do
  Else ! both inputs zero or one positive and one zero
    C = 1
  End If
Else ! at least one negative input
  C = -1 ! error
End If
F = C
return
end function N_c_m
!*****
end module CVTQTJmod

```

## Rosenbaum's Test of Location

## Main program for printing tables

```

! *****
! program:   CVRBTJ
! source:   CVRBTJ.f90
! based on: CVRBT.f90 as of 29 Apr 2002 15:22 EDT
! author:   Bruce R. Fay
! date:    18 Oct 2002 18:13 EDT
! purpose:  Generate and print critical value table for Rosenbaum's test
!           of location for 2 independent samples.
! *****
program CVRBT
use CVRBjmod
implicit none
! DECLARE VARIABLES
integer :: i, j, LU, ios, testnum
integer, dimension(:) :: CVi(4)
real*8, dimension(:) :: PVr(4)
! DEFINE FORMATS FOR OUTPUT FILE
100 format(" 1-tailed CVs at stated alpha levels")
200 format("          | - - - - - CV - - - - - | &
&- - - - - PV - - - - - |")
300 format(" n1 n2 - .05 - - .025- - .01 - - .005-      &
&- .05 - - .025- - .01 - - .005-")
400 format(2I3,4I8,4x,4F8.4)
! GET USER INPUTS
write(*,*) "Program CVRBTJ.exe by Bruce R. Fay"
write(*,*) "Generate best conservative critical values and associated"
write(*,*) "probabilities for Rosenbaum's Test for two-independent-samples"
write(*,*) "and output results to file"
write(*,*)
write(*,*) "Select one of the following:"
write(*,*)
write(*,*) " 0 - to exit program"
write(*,*) " 1 - to generate values"
write(*,*)
Do
  read(*,*) testnum
  If ( (testnum >= 0).and.(testnum <= 1) ) then
    EXIT
  Else
    write(*,*) "enter 0 - 4 please"
  End if
End Do
If (testnum == 0) GOTO 9999 ! check for user termination
! OPEN OUTPUT FILE AND WRITE FILE HEADER
LU = 8
open(unit=LU, file='CVRBTJ_.txt', iostat=ios)
IF (ios > 0 ) then
  write(*,*) "Error opening file 'CVRBTJ_.txt' "
  GOTO 9999
End if
write(LU,*) "Program CVRBTJ.exe by (Author's name here)"
write(LU,*) "Output file CVRBTJ_.txt"
write(LU,*)

```



```

write(LU,*) "Generate best conservative critical values and associated"
write(LU,*) "probabilities for Rosenbaum's Test for two-independent-samples"
write(LU,*) "based on formula in Neave & Worthington (1988)"
write(LU,*) "Distribution-free Tests, p. 148"
write(LU,*)
write(LU,*) "n1 = m, n2 = n, n1 is the size of the sample from which"
write(LU,*) "the test statistic is calculated (length of extreme run)"
write(LU,*)
! GENERATE VALUES AND OUTPUT TO FILE
call CV_RBJ_init
write(LU,100) ! print header information
write(LU,*)
write(LU,200) ! print column headers for this format
write(LU,300)
write(LU,*)
Do i = 1,50
  Do j = 1,50
    call CV_RBJbc(i,j,CVi,PVr)
    write(LU,400) i,j,CVi(1:4),PVr(1:4)
  End Do
  write(LU,*)
End Do
! CLOSE FILE
close(unit=LU, status='keep', iostat=ios)
If (ios > 0) then
  write(*,*) "Error closing file 'CVRBTJ_.txt' "
End If
9999 stop
end program CVRBT

```

### Module for generating critical values and probabilities

```

! *****
! module:   CVRBJmod
! source:   CVRBJmod.f90
! based on: CVRB4mod.f90 as of 20 Apr 2002 23:01 EDT and
!           Neave & Worthington (1988) Distribution-free Tests, Table J,
!           383-386 and Rosenbaum (1954) Tables for a nonparametric
!           test of location, Annals of Mathematical Statistics, Vol. 25,
!           146-150. The later tables also appear in Owen (1962)
!           Handbook of Statistical Tables, 499-503.
! author:   Bruce R. Fay
! date:     18 Oct 2002 18:12 EDT
! purpose:  Provide the critical values for Rosenbaum's Test of Location
!           for 2-independent-samples based on the method of Neave &
! desc:     Worthington (1988) p. 148 to calculate probability of a run of h
!           values from sample m out of a combined sample of N = m + n. The
!           formula is
!            $m!(N-h)!/[N!(m-h)!] = m/N \times (m-1)/(N-1) \times \dots \times (m-h+1)/(N-m+1)$ 
!           The value of h associated with the largest such probability that
!           is <= nominal alpha is the critical value for that situation.
!           Thus all CVs are BEST CONSERVATIVE with pr(CV) <= nominal alpha.
!           Creates the CVTs and PVTs on initialization and provides an
!           entry point that returns up to 4 critical values, and their
!           associated probabilities, based on the incoming values of m
!           and n. Checks are made that m and n are in the appropriate

```

```

!           ranges, 1 <= m <= n and 1 <= n <= 50.  The sample from which
!           the statistic is calculated must have sample size m.
!*****
module CVRBJmod
implicit none
private
public :: CV_RBJ_init, CV_RBJbc
contains
! *****
subroutine CV_RBJ_init
! INTERFACE
! There are no arguments for CV_RBJ_init.  The calling routine must call this
! subroutine once to build the CV and PV tables prior to calling CV_RBJbc()
! to obtain critical values and associated probabilities for specific n1, n2.
! The calling routine must declare an integer vector of length 4 and a real*8
! vector of length 4 and pass them in as arguments to receive the critical
! values and their associated probabilities.  For entry CV_RBJbc(m,n,CV,PV):
!   m  :: sample size for group from which the statistic is calculated
!   n  :: sample size for the other group
!   CV :: critical values vector (integer, length 4)
!   PV :: probability values vector (real, length 4)
! Unequal n1, n2 are supported for all n1, n2, both <= 50, where m is the
! sample size of the sample from which the statistic is calculated, i.e.,
! the sample with the global maximum.
! DESCRIPTION
! At entry CV_RBJ(), returns up to four critical values, if available, in
! vector CV(:), as follows:
!   CV(1) = 1-tailed alpha .05  (2-tailed alpha .10)
!   CV(2) = 1-tailed alpha .025 (2-tailed alpha .05)
!   CV(3) = 1-tailed alpha .01  (2-tailed alpha .02)
!   CV(4) = 1-tailed alpha .005 (2-tailed alpha .01)
! If a critical value is not available, a -1 is returned instead with
! associated probability 0.  Critical values may not be available because
! n1 and n2 are a) too small, b) too large, or c) too different.
! DECLARE DUMMY VARIABLES
integer, intent(in) :: m, n
integer, intent(in out), dimension(:) :: CV
real*8, intent(in out), dimension(:) :: PV
! DECLARE LOCAL VARIABLES
integer, dimension(50,50), save :: CVTbc1, CVTbc2, CVTbc3, CVTbc4
integer :: h, mm, nn, mn
real*8, dimension(50,50), save :: PVTbc1, PVTbc2, PVTbc3, PVTbc4
real*8 :: R, rm, T
logical :: p05, p025, p01, p005
CVTbc1 = -1 ! initialize the CV tables to -1 (indicates no valid entry)
CVTbc2 = -1
CVTbc3 = -1
CVTbc4 = -1
PVTbc1 = 0.0 ! initialize the PV tables to 0 (indicates no valid entry)
PVTbc2 = 0.0
PVTbc3 = 0.0
PVTbc4 = 0.0
Do nn = 1,50 ! generate the CV and PV tables
  Do mm = 1,50
    p05 = .false.
    p025 = .false.
    p01 = .false.

```

```

p005 = .false.
mn = mm + nn
T = real(mn)
rm = real(mm)
R = 1.0
Do h = 1, mm
  R = R * rm / T
  rm = rm - 1.0
  T = T - 1.0
  If( (R <= .05).and.(.not.p05) ) then
    CVTbc1(mm,nn) = h
    PVTbc1(mm,nn) = R
    p05 = .true.
  End If
  If( (R <= .025).and.(.not.p025) ) then
    CVTbc2(mm,nn) = h
    PVTbc2(mm,nn) = R
    p025 = .true.
  End If
  If( (R <= .01).and.(.not.p01) ) then
    CVTbc3(mm,nn) = h
    PVTbc3(mm,nn) = R
    p01 = .true.
  End If
  If( (R <= .005).and.(.not.p005) ) then
    CVTbc4(mm,nn) = h
    PVTbc4(mm,nn) = R
    p005 = .true.
  End If
  If (p05.and.p025.and.p01.and.p005) exit
End Do
End Do
End Do
return
! -----
entry CV_RBJbc(m,n,CV,PV)
! CV_RBJbc() must be called with m = sample size of group from which the
! statistic is calculated (group with global maximum value).
CV(:) = -1      ! initialize all return CVs to 'not available'
PV(:) = 0.0    ! initialize all return PVs to 'not available'
If ((m >= 1).and.(m <= 50).and.(n >= 1).and.(n <= 50)) then
  CV(1) = CVTbc1(m,n)
  CV(2) = CVTbc2(m,n)
  CV(3) = CVTbc3(m,n)
  CV(4) = CVTbc4(m,n)
  PV(1) = PVTbc1(m,n)
  PV(2) = PVTbc2(m,n)
  PV(3) = PVTbc3(m,n)
  PV(4) = PVTbc4(m,n)
End If
return
! -----
end subroutine CV_RBJ_init
! *****
end module CVRBJmod

```

## Kolmogorov-Smirnov Test of General Differences

## Main program for printing tables

```

! *****
! program:  CVKSTJ
! source:   CVKSTJ.f90
! based on: CVKST.f90 as of 29 Apr 2002 15:10 EDT
! author:   Bruce R. Fay
! date:    19 Oct 2002 10:59 EDT
! purpose:  Test harness for critical value modules for Kolmogorov-
!           Smirnov 2-independent-samples test for general differences.
! desc:    Provides user choice of printing critical values and
!           associated probability values for 2-sided tests based on ABS(Dmn)
!           or for 1-sided tests based on either on Dneg or Dpos.  Module
!           CVKSJmod generates the 2-sided values.
! *****
program CVKSTJ
use CVKSJmod
implicit none
! DECLARE VARIABLES
integer :: i, j, k, LU, ios, testnum
integer, dimension(:) :: CVi(4)
real, dimension(:) :: PVr(4)
! GET USER INPUTS
write(*,*) "Program CVKSTJ.exe by Bruce R. Fay"
write(*,*) "Kolmogorov-Smirnov test of general differences for"
write(*,*) "two independent samples - critical value tables with"
write(*,*) "probabilities"
write(*,*)
write(*,*) "Select one of the following:"
write(*,*)
write(*,*) " 0 - exit"
write(*,*) " 1 - generate 1-tailed CVs and actual p values using CVKSJmod"
write(*,*) " 2 - generate 2-tailed CVs and actual p values using CVKSJmod"
write(*,*)
Do
  read(*,*) testnum
  If ( (0 <= testnum).and.(testnum <= 2) ) EXIT
  write(*,*) "enter 0 - 2 please"
End Do
If (testnum == 0) GOTO 9999 ! check for user termination
! OPEN OUTPUT FILE AND WRITE FILE HEADER
LU = 8
open(unit=LU, file='CVKSTJ_.txt', iostat=ios)
IF (ios > 0 ) then
  write(*,*) "Error opening file 'CVKSTJ_.txt' "
  GOTO 9999
End if
write(LU,*) "Program CVKSTJ by (Author's name goes here)"
write(LU,*) "File CVKSTJ_.txt"
write(LU,*)
! DEFINE FORMATS FOR OUTPUT FILE
100 format(" 2-tailed CVs and PVs at stated alpha levels")
110 format(" 1-tailed CVs and PVs at stated alpha levels")
120 format("      ---- nominal alpha 2-tailed ---  &

```

```

&----- actual 2-tailed prob -----")
130 format("  n1 n2 - .10 - - .05 - - .02 - - .01 - &
          & -- .10 -- -- .05 -- -- .02 -- -- .01 --")
140 format("          ---- nominal alpha 1-tailed --- &
          &----- actual 1-tailed prob -----")
150 format("  n1 n2 - .05 - - .025 - .01 - - .005 &
          & -- .05 -- -- .025 - -- .01 -- -- .005 -")
160 format(1x,2I3,4I8,2x,4F10.6)
Select Case(testnum)
Case(1)
  write(*,*) "Outputting CVT to file for K-S 2-i-s t-g-d"
  write(*,*) "generated CVs based on Kim & Jennrich, with"
  write(*,*) "actual 1-tailed probabilities"
  write(*,*)
  write(LU,*) "Kolmogorov-Smirnov test of general differences for"
  write(LU,*) "two independent samples, critical values based on"
  write(LU,*) "Kim & Jennrich (1970,1973), with actual 1-tailed"
  write(LU,*) "probabilities generated by CVKSJmod"
  write(LU,*)
  write(*,*) "Generating CV tables"
  call CV_KSJ_init
  write(*,*) "CV_KSJ_init completed - CV tables built"
  write(LU,110) ! print header information
  write(LU,*)
  write(LU,140) ! print column headers for this format
  write(LU,150)
  write(LU,*)
  Do j = 1,50
    Do i = 1,j
      call CV_KSJbc(i,j,CVi,PVr)
      PVr = PVr/2.0
      write(LU,160) i,j,CVi(1:4),PVr(1:4)
    End Do
    write(LU,*)
  End Do
Case(2) ! 2-sided values w/ actual probabilities
  write(*,*) "Outputting CVT to file for K-S 2-i-s t-g-d"
  write(*,*) "generated CVs based on Kim & Jennrich, with"
  write(*,*) "actual 2-tailed probabilities"
  write(*,*)
  write(LU,*) "Kolmogorov-Smirnov test of general differences for"
  write(LU,*) "two independent samples, critical values based on"
  write(LU,*) "Kim & Jennrich (1970,1973), with actual 2-tailed"
  write(LU,*) "probabilities generated by CVKSJmod"
  write(LU,*)
  write(*,*) "Generating CV tables"
  call CV_KSJ_init
  write(*,*) "CV_KSJ_init completed - CV tables built"
  write(LU,100) ! print header information
  write(LU,*)
  write(LU,120) ! print column headers for this format
  write(LU,130)
  write(LU,*)
  Do j = 1,50
    Do i = 1,j
      call CV_KSJbc(i,j,CVi,PVr)
      write(LU,160) i,j,CVi(1:4),PVr(1:4)
    End Do
  End Do

```

```

        End Do
        write(LU,*)
    End Do
End Select
! CLOSE FILE
close(unit=LU, status='keep', iostat=ios)
If (ios > 0) then
    write(*,*) "Error closing file 'CVKSTJ_.txt' "
End If
9999 stop
end program CVKSTJ

```

### Module for generating critical values and probabilities

```

! *****
! module:    CVKSJmod
! source:    CVKSJmod.f90
! based on:  CVKS3mod as of 05 Jun 2002 19:00, which is based on the
!            Kim & Jennrich Tables of the exact sampling distribution of
!            the two-sample Kolmogorov-Smirnov criterion, Dmn,  $m \leq n$  in
!            Selected Tables in Mathematical Statistics, Vol. 1, 77-170
!            (1970) Harter & Owens (eds) 2nd printing (1973) with revisions,
!            published by American Mathematical Society for the
!            Institute of Mathematical Statistics
! author:    Bruce R. Fay
! date:      19 Oct 2002 10:48 EDT
! purpose:   Provide the best conservative critical values for the
!            Kolmogorov-Smirnov 2-independent-samples test for general
!            differences.
! desc:      Generates the CVTs on initialization and provides an entry
!            point that returns up to 4 critical values based on the
!            incoming values of m and n. Checks are made that
!             $1 \leq m \leq n \leq 50$ . If n1, n2 are not in this range and
!            relationship, the lookup is not performed. When CVs are
!            not available, a value of -1 is returned.
! *****
module CVKSJmod
implicit none
private
public :: CV_KSJ_init, CV_KSJbc
contains
! *****
subroutine CV_KSJ_init
! INTERFACE
! There are no arguments for CV_KSJ_init. The calling routine must call this
! subroutine once to build the CV table prior to calling CV_KSJbc() to obtain
! critical values and probabilities for specific m and n. The calling
! routine must also declare an integer vector of length 4 and pass it in to
! receive the critical values as well as a real vector of length 4 and pass
! it in to receive the probabilities. For entry CV_KSJbc(m,n,CV,PV):
!   m   :: sample size for 1st group ( $\leq n$ )
!   n   :: sample size for 2nd group
!   CV  :: critical values vector (length 4)
!   PV  :: probability values vector (length 4)
! DESCRIPTION
! At entry CV_KSJbc(m,n,CV,PV), for  $m \leq n$ , returns up to four critical

```

```

! values, if available, in vector CV(:), with actual probabilities in PV(:),
! as follows:
!   CV(1) = 1-tailed alpha .05  (2-tailed alpha .10)
!   CV(2) = 1-tailed alpha .025 (2-tailed alpha .05)
!   CV(3) = 1-tailed alpha .01  (2-tailed alpha .02)
!   CV(4) = 1-tailed alpha .005 (2-tailed alpha .01)
!   PV(1) = 1-tailed .05  (2-tailed .10) actual probability
!   PV(2) = 1-tailed .025 (2-tailed .05) actual probability
!   PV(3) = 1-tailed .01  (2-tailed .02) actual probability
!   PV(4) = 1-tailed .005 (2-tailed .01) actual probability
! If a critical value is not available, a -1 is returned instead with p = 0.0
! DECLARE DUMMY VARIABLES
integer, intent(in) :: m, n
integer, intent(out), dimension(:) :: CV
real, intent(out), dimension(:) :: PV
! DECLARE LOCAL VARIABLES
integer, dimension(50,50), save :: CVTbc10, CVTbc05, CVTbc02, CVTbc01
integer :: c, i, ixj, j
real*8, dimension(50,50), save :: PVTbc10, PVTbc05, PVTbc02, PVTbc01
real*8 :: d, pc, prevc
real*8, parameter :: p90=.90, p95=.95, p98=.98, p99=.99
logical :: f10, f05, f02, f01
CVTbc10 = -1 ! initialize CV tables to -1 (indicates no valid entry)
CVTbc05 = -1
CVTbc02 = -1
CVTbc01 = -1
PVTbc10 = 0.0 ! initialize PV tables to zero
PVTbc05 = 0.0
PVTbc02 = 0.0
PVTbc01 = 0.0
! BUILD THE CV AND PV TABLES
Do j = 1,50 ! this is n
  Do i = 1,j ! this is m
    f10 = .false.
    f05 = .false.
    f02 = .false.
    f01 = .false.
    prevc = 0.0
    ixj = i*j
    Do c = 1,ixj ! possible critical values
      d = real(c)/real(ixj) ! Dmn
      pc = akscdf(i,j,d) ! get the probability of Dmn <= C/(m*n)
      If ((.not.f10).and.(prevc >= p90).and.(pc > prevc)) then
        CVTbc10(i,j) = c
        PVTbc10(i,j) = 1.0 - prevc
        f10 = .true.
      End If
      If ((.not.f05).and.(prevc >= p95).and.(pc > prevc)) then
        CVTbc05(i,j) = c
        PVTbc05(i,j) = 1.0 - prevc
        f05 = .true.
      End If
      If ((.not.f02).and.(prevc >= p98).and.(pc > prevc)) then
        CVTbc02(i,j) = c
        PVTbc02(i,j) = 1.0 - prevc
        f02 = .true.
      End If
    End Do
  End Do
End Do

```

```

      If ((.not.f01).and.(prevc >= p99).and.(pc > prevc)) then
        CVTbc01(i,j) = c
        PVTbc01(i,j) = 1.0 - prevc
        f01 = .true.
      End If
      prevc = pc
      If ( f10.and.f05.and.f02.and.f01 ) exit
    End Do
  End Do
End Do
return
! -----
entry CV_KSJbc(m,n,CV,PV)
CV = -1      ! initialize all return CVs to 'not available'
PV = 0.0     ! initialize all probabilities to zero

If ((1 <= n).and.(n <= 50).and.(1 <= m).and.(m <= n)) then
  CV(1) = CVTbc10(m,n)
  CV(2) = CVTbc05(m,n)
  CV(3) = CVTbc02(m,n)
  CV(4) = CVTbc01(m,n)
  PV(1) = PVTbc10(m,n)
  PV(2) = PVTbc05(m,n)
  PV(3) = PVTbc02(m,n)
  PV(4) = PVTbc01(m,n)
End If
return
! -----
end subroutine CV_KSJ_init
! *****
real*8 function akscdf(a,b,d)
! From Kim & Jennrich tables of the exact sampling distribution of
! the two-sample Kolmogorov=Smirnov criterion, Dmn, m<=n in
! Selected Tables in Mathematical Statistics, Vol. 1, 77-170
! (1970) Harter & Owens (eds) 2nd printing (1973) with revisions,
! published by American Mathematical Society for the Institute of
! Mathematical Statistics.
! requires a <= b
! DECLARE DUMMY VARIABLES
integer, intent(in) :: a, b
real*8, intent(in) :: d
! DECLARE LOCAL VARIABLES
integer :: i, j
real*8 :: k, w
real*8, allocatable, dimension(:) :: u
allocate(u(b+1))
k = (real(a*b))*d + .5
u(1) = 1.
Do j = 1,b
  u(j+1) = 1.
  If (real(a*j) > k) then
    u(j+1) = 0.
  End If
End Do
Do i = 1,a
  w = real(i)/real(i+b)
  u(1) = w*u(1)

```



```

      If (real(b*i) > k) then
        u(1) = 0.
      End If
      Do j = 1,b
        u(j+1) = u(j) + (u(j+1)*w)
        If (real(IABS(b*i-a*j)) > k) then
          u(j+1) = 0.
        End If
      End Do
      End Do
      akscdf = u(b+1)
      deallocate(u)
      return
    end function akscdf
! *****
end module CVKSJmod

```

## Wilcoxon Rank-sum Test

### Main program for printing tables

```

! *****
! program:  CVWRSTJ.exe
! source:   CVWRSTJ.f90
! author:   Bruce R. Fay
! date:     25 Oct 2002  14:22 EDT
! based on: CVWRST.f90 as of 08 Jun 2002 13:02 EDT
! purpose:  Test harness for critical value tables (CVTs) for the
!           Wilcoxon rank sum test for 2-i-s.
! desc:     Provides user choice of critical value module and then
!           outputs results to a file.
! *****
program CVWRSTJ
use CVWRSJ4mod
implicit none
! DECLARE VARIABLES
integer :: i, j, LU, ios, testnum
integer, dimension(:) :: CVi(4)
real*8, dimension(:) :: PVr(4)
! GET USER INPUTS
write(*,*) "Program CVWRSTJ.exe by Bruce R. Fay"
write(*,*)
write(*,*) "Wilcoxon rank-sum test for two independent samples."
write(*,*) "Best-conservative critical values generated by method of"
write(*,*) "Wilcoxon, Katti & Wilcox (1963,68,70,73)."

```

```

If (testnum==0) GOTO 9999 ! check for user termination
! OPEN FILE FOR OUTPUT AND WRITE HEADER
LU = 8
open(unit=LU, file='CVWRSTJ_.txt', iostat=ios)
IF (ios > 0 ) then
  write(*,*) "Error opening file 'CVWRSTJ_.txt' "
  GOTO 9999
End if
write(LU,*) "File CVWRSTJ_.txt for program CVWRSTJ.exe"
write(LU,*) "by Bruce R. Fay"
write(LU,*)
write(LU,*) "Wilcoxon rank-sum test for two independent samples."
write(LU,*) "Best-conservative critical values generated by method of"
write(LU,*) "Wilcoxon, Katti & Wilcox (1963,68,70,73)."

```

## Module for generating critical values and probabilities

```

! *****
! module:    CVWRS4Jmod
! source:    CVWRS4Jmod.f90
! author:    Bruce R. Fay
! date:      25 Oct 2002 14:04
! based on:  CVWRS4mod.f90 as of 08 Jun 2002 12:52 EDT
!           Wilcoxon, Katti & Wilcox (1963) Critical values and
!           probability levels for the Wilcoxon rank sum test (and the
!           Wilcoxon signed rank test), revised Oct 1968, as it appears
!           in Harter & Owen, editors (1970,73) Selected Tables in
!           Mathematical Statistics, Volume I, 171-259.
!           (Values for  $n_1 = 1$  and  $n_1 = 2$  from Bradley (1968)
!           Distribution-free Statistical Tests, 318, Table III.)
! purpose:   Provide the BEST CONSERVATIVE 1-tailed critical values and
!           associated actual probabilities for the Wilcoxon rank sum
!           test.
! desc:      Generates CV and PV tables on initialization and provides an
!           entry point that returns up to 4 critical values based on the
!           incoming values of  $n_1$  and  $n_2$ . Checks are made that  $n_1$ ,  $n_2$ 
!           are in the appropriate range and relationship for the tables,
!           with  $1 \leq n_1 \leq n_2 \leq 50$ .
! *****
module CVWRSJ4mod
implicit none
private
public :: CV_WRSJ4_init, CV_WRSJ4bc
contains
! *****
subroutine CV_WRSJ4_init
! INTERFACE
! There are no arguments for CV_WRSJ4_init. The calling routine must call
! this subroutine once to build the CV table prior to calling CV_WRSJ4bc() to
! obtain critical values for specific  $m$  and  $n$ . The calling routine must
! declare two vectors and pass them as arguments: an integer vector of length
! 4 to receive the critical values and a real*8 vector of length 4 to receive
! the associated probabilities. For entry CV_WRSJ4bc(a,b,CV,PV):
!   a  :: sample size for 1st group ( $\leq b$ )
!   b  :: sample size for 2nd group
!   CV :: critical values vector (length 4)
!   PV :: actual probability values vector (length 4)
! DECLARE DUMMY VARIABLES
integer, intent(in) :: a, b
integer, intent(out), dimension(:) :: CV
real*8, intent(out), dimension(:) :: PV
! DECLARE LOCAL VARIABLES
integer :: h, i, j, k, k1, k2, M, minRS, N, RS, u, ub
integer, dimension(50,50), save :: CVTbc10, CVTbc05, CVTbc02, CVTbc01
real*8, dimension(50,50), save :: PVTbc10, PVTbc05, PVTbc02, PVTbc01
real*8, allocatable, dimension(:) :: cf, f, z
real*8 :: Pr, Prev
real*8, parameter :: p05=0.05, p025=0.025, p01=0.01, p005=0.005
real*8, parameter :: oneppt = 0.001
logical :: f10, f05, f02, f01, Pr_underflow, Prev_underflow

```

```

CVTbc10 = -1  ! initialize CV and PV tables
CVTbc05 = -1
CVTbc02 = -1
CVTbc01 = -1
PVTbc10 = 0.
PVTbc05 = 0.
PVTbc02 = 0.
PVTbc01 = 0.
Do N = 2,50
  Do M = 1,N  ! build the z vector
    minRS = M*(M+1)/2
    k = (M+50)**2
    allocate (z(0:k))
    z = 0.
    Do i = 1,N
      Do j = 1,k
        k1 = (M+i)*j - 1
        K2 = i*j - 1
        If (k1 <= k) then
          z(k1) = z(k1) - real(M+i)
        End If
        If (k2 <= k) then
          z(k2) = z(k2) + real(i)
        End If
        If (k1 > k .and. k2 > k) exit
      End Do
    End Do
  ! build the freq and cumfreq vector and find the critical values
  f10 = .false.
  f05 = .false.
  f02 = .false.
  f01 = .false.
  ub = (M+N)*(M+N+1)/2      ! set upper bound on u
  allocate (f(0:ub))      ! allocate the frequency vector
  allocate (cf(0:ub))      ! and the cumulative frequency vector
  f = 0.
  f(0) = 1.
  cf = 0.
  cf(0) = 1.
  Do u = 1,ub
    Do h = 0,(u-1)
      f(u) = f(u) + ( f(h)*z(u-h-1) )
    End Do
    f(u) = f(u)/u
    cf(u) = cf(u-1) + f(u)
    Pr = cf(u)
    Prev = cf(u-1)
    Pr_underflow = .false.
    Prev_underflow = .false.
  ! The probabilities Pr and Prev get smaller with each pass
  ! through the following loop. Thus, once they both drop below
  ! oneppt (see declaration) there is no point continuing the loop.
  Do i = 1,M
    If (Pr > oneppt) then
      Pr = Pr*(M+1-i)/(N+i)
    Else
      Pr_underflow = .true.
    End If
  End Do
End Do

```

```

      End If
      If (Prev > oneppt) then
        Prev = prev*(M+1-i)/(N+i)
      Else
        Prev_underflow = .true.
      End If
      If (Pr_underflow .AND. Prev_underflow) exit
    End Do
    RS = minRS + u-1 ! rank sum = M(M+1)/2 + u-1
! Find the best conservative CVs for specified alphas
    If ((Prev <= p05).and.(Pr > p05).and.(.not.f10)) then
      CVTbc10(M,N) = RS
      PVTbc10(M,N) = Prev
      f10 = .true.
    End If
    If ((Prev <= p025).and.(Pr > p025).and.(.not.f05)) then
      CVTbc05(M,N) = RS
      PVTbc05(M,N) = Prev
      f05 = .true.
    End If
    If ((Prev <= p01).and.(Pr > p01).and.(.not.f02)) then
      CVTbc02(M,N) = RS
      PVTbc02(M,N) = Prev
      f02 = .true.
    End If
    If ((Prev <= p005).and.(Pr > p005).and.(.not.f01)) then
      CVTbc01(M,N) = RS
      PVTbc01(M,N) = Prev
      f01 = .true.
    End If
    If (f10.and.f05.and.f02.and.f01) exit ! found all 4 CVs!
  End Do
  deallocate(z,f,cf)
End Do
return
! -----
entry CV_WRSJ4bc(a,b,CV,PV)
CV = -1 ! initialize all return CVs to 'not available'
PV = 0. ! initialize all return p's to zero
If ((b >= 1).and.(b <= 50).and.(a >= 1).and.(a <= b)) then
  CV(1) = CVTbc10(a,b)
  CV(2) = CVTbc05(a,b)
  CV(3) = CVTbc02(a,b)
  CV(4) = CVTbc01(a,b)
  PV(1) = PVTbc10(a,b)
  PV(2) = PVTbc05(a,b)
  PV(3) = PVTbc02(a,b)
  PV(4) = PVTbc01(a,b)
End If
return
! -----
end subroutine CV_WRSJ4_init
! *****
end module CVWRSJ4mod

```

## A Program for Generating All Permutations of $\{1, 2, \dots, n\}$

Robert DiSario  
Bryant College

---

A Visual Basic program that generates all permutations of  $\{1, 2, \dots, n\}$  is presented. The procedure for running the program as an Excel macro is described. An application is presented which involves selecting permutations which meet a specific constraint.

Key words: Visual Basic, permutation.

---

### Introduction

A Visual Basic program for generating all combinations of  $n$  elements taken  $m$  at a time was presented in Stamatopoulos (2002). The present work presents a program for generating all permutations of  $n$  elements. Applications involving combinations and permutations often arise in designing experiments and in other areas. As an example, the program was used to find all permutations that meet a specific requirement.

The procedure given in the present work meets the requirements stated in Stamatopoulos (2002) for algorithms which implement automatic enumeration: a) all possible cases are exhausted; b) none of the permutations need to be stored – the current case that has been formulated is the basis for generating the next one. Therefore it presents a practical means for generating permutations.

### Methodology

The program consists of a main module, `Macro1()`, and 3 functions: `Permute()`, `Findlarg()` and `Sort()`. The main module handles input and output (input from Excel ; output to a text file), dimensions and initializes an array, and calls

---

Robert DiSario is an Assistant Professor in the department of mathematics at Bryant College. He received a Ph.D. in statistics from Boston University in 1996. His academic interests include applied statistics and combinatorics. E-mail him at [rdisario@bryant.edu](mailto:rdisario@bryant.edu)

`Permute()`. `Findlarg()` returns the largest element to the right of a given position in an array. `Sort()` sorts the elements to the right of a given position in an array. `Permute()` takes as input a permutation of  $\{1, 2, \dots, n\}$  and creates the next permutation in the “natural sequence”. For an example of the “natural sequence” of permutations of  $\{1, 2, 3, 4\}$  see the output below. `Permute()` also returns 0 when the final permutation in the natural sequence has been created. A general description of `Permute()` follows. A listing of the program, written in Visual Basic, appears in an appendix.

### Description of `Permute()` function

`Permute(x(), n)`

Set `bigfix = n`.

Note: `bigfix` is an element that serves as a reference point in the array.

#### **Top:**

Find position of `bigfix` (call it `bigindx`). Check whether array is in descending order from `bigindx` to the right. If descending, work left. Else, work right.

**Work left:** (refers to left of `bigfix`)

If nothing to left of `bigfix`, then done (this is the last permutation in natural sequence).

Else the element to the left of `bigfix`, `x(bigindx-1)`, needs to be changed. Switch it with the smallest element on its right

which is bigger than it. Then sort the elements from *bigindx* to the right.

*Permute( )* is done (indicated by *done = 1*).

**end Work left**

**Work right:** (refers to right of *bigfix*)

Find the largest element on the right of *bigfix*. Set *bigfix* equal to this largest element.

*Permute( )* is not done (indicated by *done = 0*)

**end Work right**

**Return to top:**

Results

#### Application 1

As a first example, the program was used to generate all 24 permutations of the set

$\{1,2,3,4\}$ . The results are shown in Table 1. This output reveals the order referred to above as the “natural sequence”. Note that the output file contains a single column of permutations, but that Table 1 has been reformatted into 6 columns to save space.

#### Application 2

As a typical application, experimenters are often interested in the order of presentation of experimental conditions or stimuli. In some cases, the orders used must be selected according to very specific considerations. Furthermore, the experimenter may desire to use a different order for each of the subjects or replications. As an example, suppose an experimenter wants a list of all the permutations of  $\{1,2,3,4,5\}$  in which “1” is not next to “2”, “2” is not next to “3”, “3” is not next to “4”, and “4” is not next to “5”. The program was modified (as described below) to check each permutation to determine whether or not it meets this constraint. The list of all such permutations appears in Table 2.

Table 1. “Natural Sequence” of Permutations of  $\{1,2,3,4\}$ . Read down then across.

1 2 3 4	1 4 2 3	2 3 1 4	3 1 2 4	3 4 1 2	4 2 1 3
1 2 4 3	1 4 3 2	2 3 4 1	3 1 4 2	3 4 2 1	4 2 3 1
1 3 2 4	2 1 3 4	2 4 1 3	3 2 1 4	4 1 2 3	4 3 1 2
1 3 4 2	2 1 4 3	2 4 3 1	3 2 4 1	4 1 3 2	4 3 2 1

Table 2. All permutations of  $\{1,2,3,4,5\}$  with the property that adjacent elements are not consecutive integers.

1 3 5 2 4	2 4 1 3 5	2 5 3 1 4	3 1 5 2 4	3 5 2 4 1	4 2 5 1 3	5 2 4 1 3
1 4 2 5 3	2 4 1 5 3	3 1 4 2 5	3 5 1 4 2	4 1 3 5 2	4 2 5 3 1	5 3 1 4 2

To select only those permutations that meet the constraint, the section of the program that prints the permutation was modified. First the permutation was checked to see if it satisfies the constraint. Then printing was conditional on the outcome of this check. This was accomplished by setting a “satisfy” flag to 0 if the constraint was not met and to 1 if the constraint was met. The specific lines that were changed (both original and modified) are presented in Appendix III. A similar

approach could be used to select permutations according to other constraints.

#### References

Stamatopoulos, C. (2002). Generation of combinations using *Excel*. *Journal of Modern Applied Statistical Methods*, 1, 191-194.

## Appendix I

The BASIC code that appears in Appendix II can be run as an Excel macro. The procedure for doing this is described in Stamatopoulos (2002). Note that before pasting the program lines into the Visual Basic editor, it is necessary to first delete

two lines which are automatically generated by Excel: *Sub Macro1()* and *End Sub*.

The program can be assigned to a control key. It will read a value of  $n$  from the cell *B4* in *Sheet1* of the *Excel* workbook. It outputs the permutations to a text file called *perms.txt*.

## Appendix II

## Program listing

```
Sub Macro1()
'Open file for output.
'Read n from worksheet
'Set initial permutation {1,2,...,n}
Open "c:\perms.txt" For Output As #1
n = Range("B4")
ReDim x(n)
For i = 1 To n
  x(i) = i
Next i

'Notdun=0 iff current permutation is n, n-1, ..., 1
notdun = 1
Do While (notdun)
  For i = 1 To n
    'Print current permutation
    Print #1, x(i);
  Next i
  'Print line feed
  Print #1, ""
  'Find next permutation and note whether it is the final one
  notdun = permute(x(), n)
Loop
Close
End Sub

Function permute(x(), n)
'Creates the next permutation in the "natural sequence"
'Returns 0 if permutation is n, n-1, ..., 1
'Default is to return 1
permute = 1
bigfix = n
'Done = 1 indicates next permutation is complete, 0 not.
done = 0
Do While (done = 0)
  done = 1

'Find the index of bigfix
  For i = 1 To n
    If x(i) = bigfix Then bigindx = i
  Next i
  descend = 1
  If bigindx <> n Then
    For i = bigindx To n - 1
      If x(i) < x(i + 1) Then descend = 0
```



```

Next i
End If
If descend And bigindx = 1 Then permute = 0
If descend Then
'Work left
  current = x(bigindx - 1)
  candidx = bigindx
'Find element to switch with x(bigindx-1)
  For i = bigindx To n
    If x(i) > current And x(i) < x(candidx) Then candidx = i
  Next i
'Switch them
  temp = x(candidx)
  x(candidx) = x(bigindx - 1)
  x(bigindx - 1) = temp
  temp = sort(x(), bigindx)
End If
'End of work left

'Work right
If descend = 0 Then
  done = 0
  bigfix = findlarg(x(), bigindx + 1)
End If
'End of work right
Loop
End Function

Function findlarg(x(), start)
'Finds largest x(i) from i = start to i = n
candid = x(start)
ub = UBound(x)
For i = start To ub
  If x(i) > candid Then candid = x(i)
Next i
findlarg = candid
End Function

Function sort(x(), start)
'Sorts x() from i = start to i = n
ub = UBound(x)
For i = start To ub
  For j = i To ub
    If x(i) > x(j) Then
      temp = x(i)
      x(i) = x(j)
      x(j) = temp
    End If
  Next j
Next i

End Function

```

## Appendix III

Program modification used to select permutations meeting constraint described in application 2.

Original code:

```
For i = 1 To n
'Print current permutation
  Print #1, x(i);
Next i
'Print line feed
Print #1, ""
```

Modified code:

```
'Check whether permutation meets constraints
satisfy = 1
For i = 2 To n
If Abs(x(i) - x(i - 1)) = 1 Then satisfy = 0
Next i

If satisfy Then
  For i = 1 To n
' print current permutation
    Print #1, x(i);
  Next i
' print line feed
  Print #1, ""
End If
```

# DataMineIt<sup>SM</sup>

announces

# PermutelIt<sup>TM</sup> v2.0

## The fastest, most comprehensive and robust permutation test software on the market today.

Permutation tests increasingly are the statistical method of choice for addressing business questions and research hypotheses across a broad range of industries. Their distribution-free nature maintains test validity where many parametric tests (and even other nonparametric tests), encumbered by restrictive and often inappropriate data assumptions, fail miserably. The computational demands of permutation tests, however, have severely limited other vendors' attempts at providing useable permutation test software for anything but highly stylized situations or small datasets and few tests. PermutelIt<sup>TM</sup> addresses this unmet need by utilizing a combination of algorithms to perform two-sample, non-parametric permutation tests very quickly – often more than an order of magnitude faster than widely available commercial alternatives when one sample is large and many tests and/or multiple comparisons are being performed (which is when runtimes matter most). PermutelIt<sup>TM</sup> can make the difference between making deadlines, or missing them, since data inputs often need to be revised, resent, or re-cleaned, and one hour of runtime quickly can become 10, 20, or 30 hours.

In addition to its speed even when one sample is large, some of the unique and powerful features of PermutelIt<sup>TM</sup> include:

- the availability to the user of a wide range of test statistics for performing permutation tests on continuous, count, & binary data, including: pooled-variance t-test; separate-variance Behrens-Fisher t-test; joint tests for scale and location coefficients; Brownie et al. "modified" t-test; exact inference; Poisson normal-approximate test; Fisher's exact test
- extremely fast exact inference (no confidence intervals – just exact p-values) for most count data and high-frequency continuous data
- the availability to the user of a wide range of multiple testing procedures, including: Bonferroni, Sidak, Stepdown Bonferroni, Stepdown Sidak, Stepdown Bonferroni and Stepdown Sidak for discrete distributions, Hochberg Stepup, FDR, Dunnett's one-step (for MCC under ANOVA assumptions), Stepdown Permutation (for FWE, FDR, and FDP), Permutation-style adjustment of permutation p-values
- fast, efficient, and automatic generation of all pairwise comparisons
- efficient variance-reduction under conventional Monte Carlo via self-adjusting permutation sampling when confidence intervals contain the user-specified critical value of the test
- maximum power under conventional Monte Carlo via a new sampling optimization technique (see Opdyke, JMASM, Vol. 2, No. 1: forthcoming, May, 2003)
- fast permutation-style p-value adjustments for multiple comparisons (the code is designed to provide an additional speed premium for these resampling-based multiple testing procedures)
- simultaneous permutation testing and permutation-style p-value adjustment, although for relatively few tests at a time (this capability is not even provided as a preprogrammed option with any other software currently on the market)

For Telecommunications, Pharmaceuticals, fMRI data, Financial Services, Clinical Trials, Insurance, Bioinformatics, and just about any data rich industry where large numbers of distributional null hypotheses need to be tested on samples that are not extremely small and parametric assumptions are either uncertain or inappropriate, PermutelIt<sup>TM</sup> is the optimal, and only, solution.

To learn more about how PermutelIt<sup>TM</sup> can be used for your enterprise, and to obtain a demo version in early 2003, please contact its author, J.D. Opdyke, President, DataMineIt<sup>SM</sup>, at [jdopdyke@datamineit.com](mailto:jdopdyke@datamineit.com) or [www.datamineit.com](http://www.datamineit.com).

DataMineIt<sup>SM</sup> is a technical consultancy providing statistical data mining, econometric analysis, and data warehousing services and expertise to the industry, consulting, and research sectors. PermutelIt<sup>TM</sup> is its flagship product.

# Announcing StatXact 5!

StatXact 5, with over 100 procedures and a 1500 page manual that is really a textbook on exact methods, provides the world's most comprehensive collection of exact procedures for significance tests and confidence intervals. Among its new features, StatXact 5 now gives you a host of new procedures for the commonly-encountered two-binomial situation. Based on recent research (Agresti and Min, *Biometrics* 2000; Chan and Zang, *Biometrics* 1999), these procedures will give you more powerful exact p-values, and shorter exact confidence intervals.

## New In StatXact 5

### Unconditional exact tests for 2 binomials:

- Superiority
- Non-inferiority
- Equivalence

More powerful exact unconditional tests and shorter exact confidence intervals for differences and ratios of proportions

Unconditional exact McNemar's test

Exact interaction tests in stratified 2xC tables

- comparison of C ordered binomials
- comparison two ordered multinomials

Exact test of trend for correlated binary data

Exact tests and confidence intervals for stratified Poisson data

While some standard software programs have a few exact tests, none has anywhere near the coverage of StatXact 5. StatXact 5, with over 100 tests and procedures, gives you exact p-values and confidence intervals for one-, two- and k-sample problems,  $R \times C$  contingency tables, stratified  $2 \times 2$  and  $2 \times C$  contingency tables, goodness-of-fit tests, measures of association, binomial data, multinomial data, and censored survival data. Plus, StatXact 5 gives you exact power and sample size capabilities.

CYTEL Software Corporation • 675 Massachusetts Ave., Cambridge, MA 02139 USA  
Tel (617) 661-2011 • Toll Free (US) 866-298-3511 • Fax (617) 661-4405  
<http://www.cytel.com> • E-mail: [sales@cytel.com](mailto:sales@cytel.com)

INTERNATIONAL DISTRIBUTORS: Ask Int'l (UK) e-mail: [cyteluk@asru.com](mailto:cyteluk@asru.com) • Tel: +44(0) 1227 795 240 • Fax: +44(0) 1227 795 201; ID2 (Belgium)  
• Tel: 32 2 6468918 • Fax: 32 2 4468662; Spadille Biostatistics ApS (Denmark) • [spadille@spadille.dk](mailto:spadille@spadille.dk) • Tel: 4548.484100 • Fax: 4248484200

**Cytel**  
STATISTICAL SOFTWARE

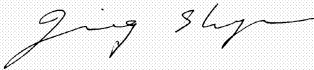
# Depth and flexibility for predicting numerical outcomes

As a clear leader in statistical software, SPSS has what you need for analysis — and the complete analytical process.

SPSS is a modular, tightly integrated, full-featured product line. It's available for Windows and Macintosh desktops. Alternatively, it's available for many high-performance server platforms. The SPSS product line covers the full analytical process. SPSS' offering includes products for database access, data cleaning and management, as well as a broad range of analytical capabilities, and high-quality tabular and graphical output. You can even publish your SPSS results to the Web. This enables people who don't have SPSS installed on their machines to interact with results using their Web browsers. SPSS products are available through a variety of pricing and licensing programs, including student, graduate student and campus-wide licenses.

Take a look at some highlights in SPSS' line-up for predicting numerical outcomes and learn about just one aspect of SPSS' many offerings for the analytical process.

Sincerely,



Jing Shyr, Ph.D.  
Vice President and Chief Statistician  
SPSS Inc.



Kyle A. Weeks, Ph.D.  
Senior Product Manager  
SPSS Inc.

## Linear Mixed Models procedure

Do you have data that display correlation and non-constant variability, such as data that represent students nested within classrooms or consumers nested within families? You can model not only means but also variances and covariances in your data using the powerful Linear Mixed Models procedure. Its flexibility means you can formulate a wide variety of models, such as multilevel models with fixed-effects covariances, hierarchical-linear models, random-effects models, random-coefficient models and linear-growth models. In addition, you can work with repeated measure designs, including incomplete repeated measurements in which the number of observations varies across subjects.

## General Linear Models (GLM) procedure — multivariate

Do you need a flexible procedure that works simultaneously with related multiple dependent variables? SPSS' GLM multivariate procedure does just that — providing flexible design and contrast options to estimate means and variances and to test and predict means. Mix and match categorical and continuous data to build models. Because GLM multivariate doesn't limit you to one data type, you have options giving you a wealth of model-building possibilities. Also, you can easily visualize relationships using profile plots (interaction plots) resulting from estimated predicted mean values.

## General Linear Models (GLM) procedure — repeated measures

Do you need to measure the same people over time, for example, to measure how overall employee satisfaction increases or decreases? Using SPSS' GLM repeated measures procedure you can analyze variances when you make the same measurement a fixed number of times on

individual subjects or cases. Get the flexibility to mix and match categorical and continuous-level predictors — including interactions. As with the GLM multivariate procedure, you can see relationships in your data using profile plots.

## Nonlinear Regression (NLR) and Constrained Nonlinear Regression (CNLR) procedures

Are you working with models that have nonlinear relationships, such as predicting coupon redemption as a function of time and number of coupons distributed? Estimate nonlinear equations using one of two SPSS procedures: NLR for unconstrained problems and CNLR for both constrained and unconstrained problems. CNLR empowers you to write your own algorithms. CNLR also gives you the flexibility to:

- Use linear and nonlinear constraints on any combination of parameters
- Estimate parameters by minimizing any smooth loss function (objective function)
- Compute bootstrap estimates of parameter standard errors and correlations

## Everything you need for predicting numerical outcomes

SPSS' procedures for predicting numerical outcomes aren't limited to the ones we just described. The following procedures help give SPSS 11.0 what you need for prediction:

- Linear Regression
- Weighted Least Squares Regression
- Two-Stage Least Squares
- Survival Analysis procedures
  - Cox Regression with time-dependent covariates
  - Kaplan-Meier
  - Life Tables

Want to know what other statistics — including stats for identifying groups and time-series analysis — and software SPSS offers for the complete analytical process? Visit [www.spss.com/statisticalmethods](http://www.spss.com/statisticalmethods) to download a white paper, "Complete end-to-end analysis with SPSS 11.0." Do you like what you see? You can buy SPSS 11.0 online at [www.spss.com/store](http://www.spss.com/store) or call (800) 543-9247.

SPSS BI helps people solve business problems using statistics and data mining. This predictive technology enables our customers in the commercial and public sectors to make better decisions and improve results. SPSS BI software and services are used successfully in a wide range of applications, including customer attraction and retention, cross-selling, survey research, fraud detection, Web site performance, forecasting and scientific research. SPSS BI's market-leading products include SPSS® Clementine®, AnswerTree®, DecisionTime® and SigmaPlot®

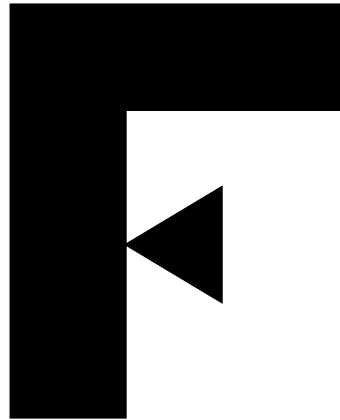


Call today  
for an SPSS  
product catalog  
(800) 543-9247

*“Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away.”*

- Antoine de Saint Exupery

F is a carefully crafted subset of the most recent version of Fortran, the world’s most powerful numeric language.



Using F has some very significant advantages:

- Programs written in F will compile with any Fortran compiler
- F is easier to use than other popular programming languages
- *F compilers are free* and available for Linux, Windows, and Solaris
- Several books on F are available
- F programs may be linked with C, Fortran 95, or older Fortran 77 programs

F retains the modern features of Fortran—modules and data abstraction, for example—but discards older error-prone facilities of Fortran.

It is a safe and portable programming language.

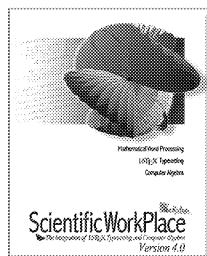
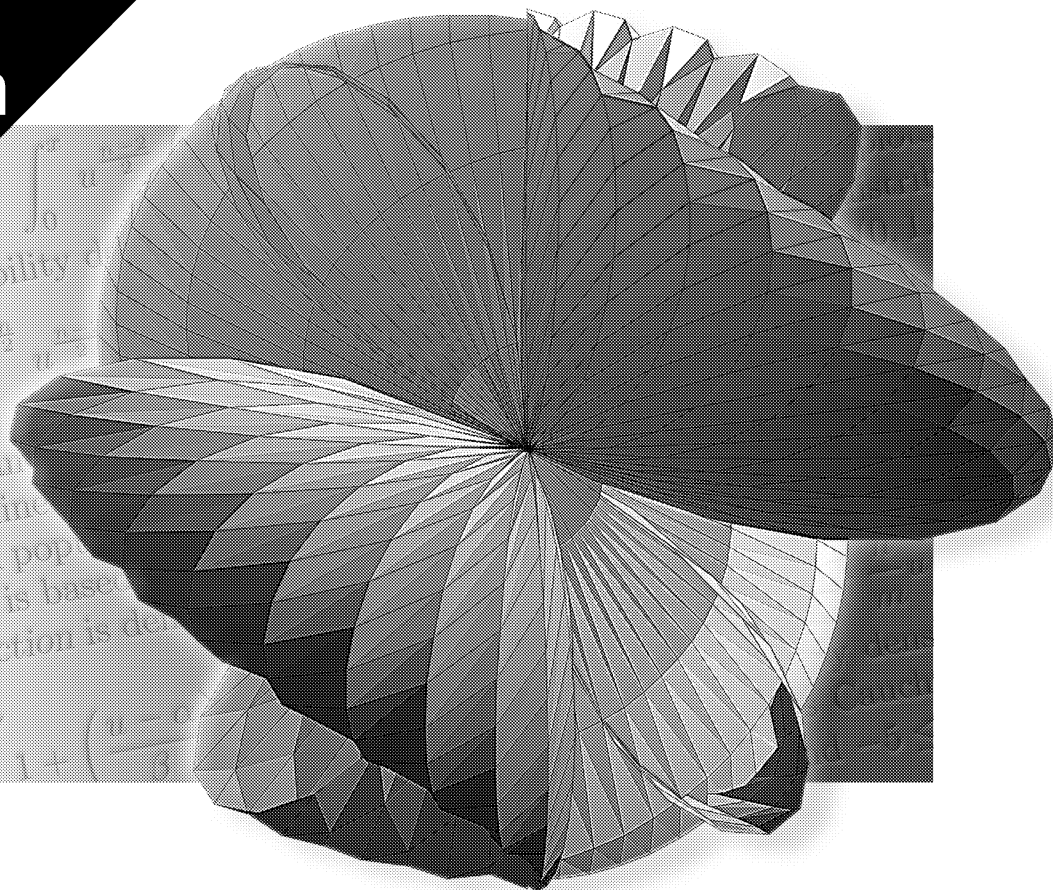
F encourages Module-Oriented Programming.

It is ideal for teaching a programming language in science, engineering, mathematics, and finance.

It is ideal for new numerically intensive programs.

The Fortran Company  
11155 E. Mountain Gate Place, Tucson, AZ 85749 USA  
+1-520-256-1455 +1-520-760-1397 (fax)  
<http://www.fortran.com> [info@fortran.com](mailto:info@fortran.com)

Introducing  
**Version  
4.0**



Math Word Processing  
**L<sup>A</sup>T<sub>E</sub>X** Typesetting  
Computer Algebra

The Gold Standard for Mathematical Publishing  
and the Easiest-to-Use Computer Algebra System

*Now in a new version!*

*Scientific WorkPlace* makes writing, publishing, and  
doing mathematics easier than you ever imagined  
possible.

- ◆ Enter text and mathematics naturally in the same paragraph
- ◆ Produce documents with or without **L<sup>A</sup>T<sub>E</sub>X** typesetting
- ◆ Produce portable **L<sup>A</sup>T<sub>E</sub>X** output
- ◆ Perform mathematical computations with both the *MuPAD*<sup>®</sup> and *Maple*<sup>®</sup> computer algebra engines
- ◆ Export documents as HTML, with mathematics exported as graphics or as MathML
- ◆ Use hyperlinking to create an entire web of *Scientific WorkPlace* documents
- ◆ And more

**MacKichan**  
SOFTWARE, INC.

# Scientific WorkPlace<sup>®</sup>

Email: [info@mackichan.com](mailto:info@mackichan.com) ◆ Toll Free: 877-724-9673 ◆ Fax: 206-780-2857

Visit our website for free evaluation copies of all our software.

[www.mackichan.com/jmsm](http://www.mackichan.com/jmsm)





# Find Your Path With Us

**HIGHLY SPECIALIZED PERMANENT & CONTRACT  
OPPORTUNITIES AVAILABLE**

- BIostatISTICS
- SAS® PROGRAMMING
- DATA MANAGEMENT
- MARKETING SCIENCE
- RESEARCH
- MANUFACTURING

## **Permanent Placement**

Contact: Tracey Gmoser

800.989.5627

Fax: 212.818.9067

[perm@smithhanley.com](mailto:perm@smithhanley.com)

[www.smithhanley.com](http://www.smithhanley.com)

## **Contract Staffing**

Contact: Keith Shelly

800.684.9921

Fax: 407.805.3020

[contract@smithhanley.com](mailto:contract@smithhanley.com)

[www.smithhanley-consulting.com](http://www.smithhanley-consulting.com)

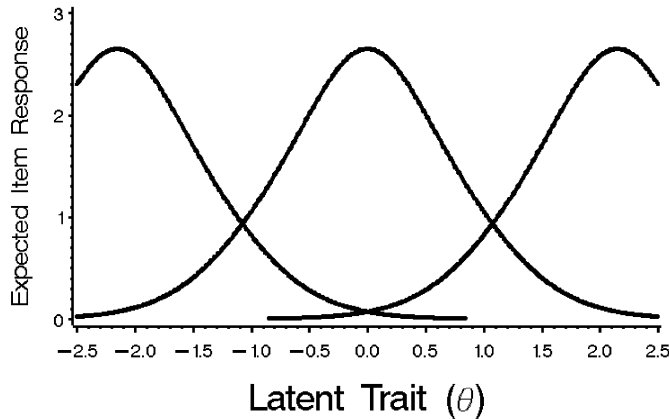
*Smith Hanley*

New York • Chicago • Houston • Southport • Orlando



# ***GGUM2000***

## *Item Response Theory Models for Unfolding*



The GGUM2000 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests

that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2000 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected a posteriori (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. The software is accompanied by a detailed user's manual. **GGUM2000 is free** and can be downloaded from:

<http://www.education.umd.edu/EDMS/tutorials>

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2000 today!

**Letitia Uduma  
Instructional Technology  
Consultant**

**Services Include:**

**Instructional Design  
Instruction Development  
Multimedia Integration  
Systems Analysis & Evaluation  
Performance Support  
E-Learning  
Computer Based Training  
Traditional Training  
Document Design**

***"For All Your IT Needs"***

**Contact Us at  
(313) 272-6532 - Phone  
(313) 272-1560 - Fax  
[letitia.uduma@davenport.edu](mailto:letitia.uduma@davenport.edu) - email**

***Committed to Excellent Service and Education***



## Are you involved in Data Modeling or Data Mining?

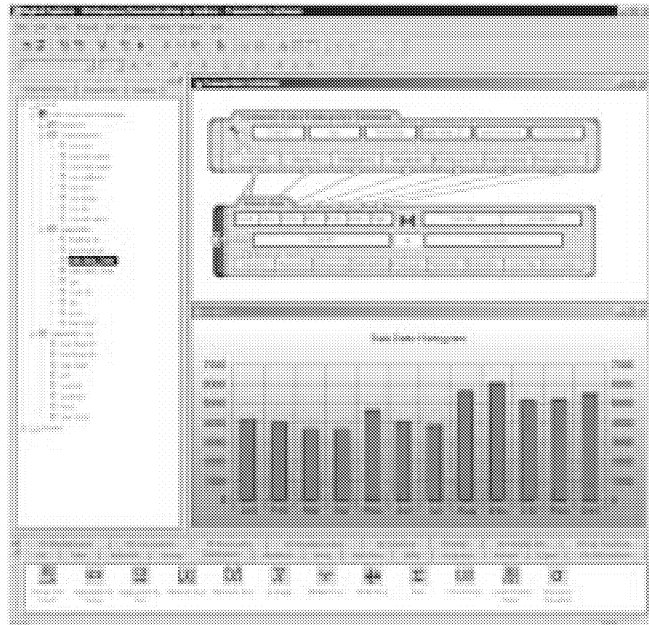
## Are you spending a large percentage of your time dealing with data issues?

If so, you will be happy to know that we have developed a tool that specifically addresses the data prep tasks associated with data modeling and data mining. The tool is called the Digital Excavator from Digital Archaeology ([www.digarch.com](http://www.digarch.com)). Data modelers are well aware of the time-consuming and sometimes frustrating nature of data set-up. In many cases data preparation can represent 60%-80% of the data mining project length. With Digital Archaeology's Digital Excavator, data preparation tasks are streamlined, results are more accurate, and the modeler has more time to focus on finding the appropriate mathematical solution--rather than wasting time with painful data issues. Digital Archaeology's software is intuitive, visual, self-documenting, and deploys what a number of analysts and customers have termed the "most elegant" user interface for data analysis and exploration ever conceived. It's the only tool specifically designed for the data prep tasks of data modeling.

**Visit our website and see for yourself! >>>> [www.digarch.com](http://www.digarch.com)**

Functions have been created which perform the following:

- Frequency Distributions
- Categorical Variable Profile
- Continuous Variable Profile
- Histograms
- De-duping
- Find and Replace Missing Values
- Find and Split Out Outliers
- Binning
- Correlation Matrix
- Cross-Tabs
- Panel Variables (Occupancy Map)
- Lag functions
- Decimal Scaling
- Rank and Sample Variables
- Recency, Frequency, Monetary Analysis
- N-Tile Distributions
- Gains Charts
- Many others



15721 COLLEGE BOULEVARD  
LENEXA, KS 66219  
1-877-DIGARCH (344-2724)  
[WWW.DIGARCH.COM](http://WWW.DIGARCH.COM)



# Ready to Take Your Next Step?

As your career climbs and  
each step requires careful  
planning, consider  
The Cambridge Group...

...Your Success is Our Business.

#### **Business Statistics**

- Quantitative Analysis
- Marketing Sciences/Research
- Econometrics
- Quality Assurance

[stat@cambridgegroup.com](mailto:stat@cambridgegroup.com)

#### **Consultant & Contract Staffing**

- Biostatisticians
- SAS®/Statistical Programmers
- Clinical Data Managers
- Clinical Systems
- CRA's & Clinical Monitors
- Medical Writers
- Project Managers
- Bioinformatics

[contract@cambridgegroup.com](mailto:contract@cambridgegroup.com)

#### **Clinical Computing & Data Management**

- SAS Programming/Application
- Clinical Data Management
- Systems Design & Analysis

[QA@cambridgegroup.com](mailto:QA@cambridgegroup.com)

#### **Biostatistics**

- Clinical
- Preclinical/Nonclinical
- Health Outcomes
- PK/PD

[biostat@cambridgegroup.com](mailto:biostat@cambridgegroup.com)

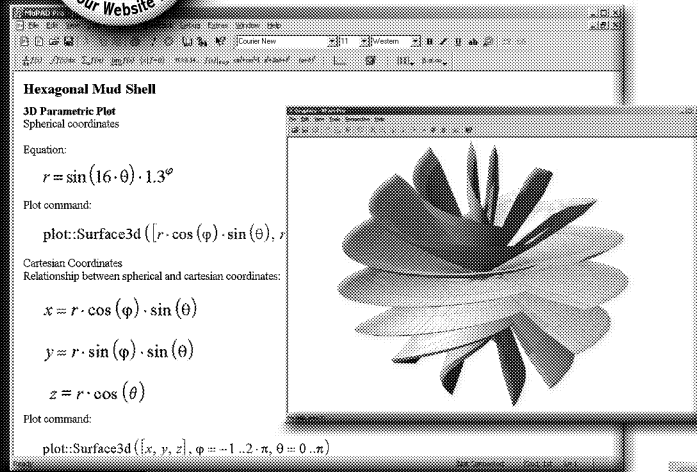


**THE CAMBRIDGE GROUP LTD**

**(800) 525-3396 fax (203) 226-3856**  
**[www.cambridgegroup.com](http://www.cambridgegroup.com)**

# MuPAD<sup>®</sup> Pro

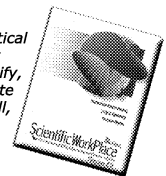
Version  
**2.0**  
See our Website for details



## The Open Computer Algebra System

MacKichan Software is proud to bring you MuPAD Pro, a full-featured computer algebra system in an integrated and open environment for symbolic and numeric computing. The MuPAD language has a Pascal-like syntax and allows imperative, functional, and object-oriented programming. Its domains and categories are like object-oriented classes that allow over-riding and overloading methods and operators, inheritance, and generic algorithms. A comfortable notebook interface includes a graphics tool for visualization, an integrated source-level debugger, a profiler, and hypertext help.

*Scientific WorkPlace, the proven solution for mathematical publishing, is an excellent companion to MuPAD Pro. It integrates with MuPAD so that you can evaluate, simplify, solve, and plot from inside your document, and evaluate functions that you have defined with MuPAD. Best of all, you typeset in LaTeX with just the click of a button.*



**MacKichan**  
SOFTWARE, INC.

Tools for Scientific Creativity Since 1981

Toll Free: 877-724-9673 • Email: [info@mackichan.com](mailto:info@mackichan.com)

Go to our homepage for free trial versions of all our products.

[www.mackichan.com/jmsm](http://www.mackichan.com/jmsm)

## Announcing the highly-anticipated new Numerical Recipes products

### Numerical Recipes in C++

The Art of Scientific Computing  
Second Edition

William H. Press, Saul A. Teukolsky,  
William T. Vetterling, and  
Brian P. Flannery

"This monumental and classic work is beautifully produced and of literary as well as mathematical quality. It is an essential component of any serious scientific or engineering library."

—*Computing Reviews*

This new version incorporates completely new C++ versions of the more than 300 *Numerical Recipes Second Edition* routines widely recognized as the most accessible and practical basis for scientific computing, in addition to including the full mathematical and explanatory contents of *Numerical Recipes in C*.

### Key Features:

- Includes linear algebra, interpolation, special functions, random numbers, nonlinear sets of equations, optimization, eigensystems, Fourier methods and wavelets, statistical tests, ODEs and PDEs, integral equations, and inverse theory.
- The routines, in ANSI/ISO C++ source code, can be used with almost any existing C++ vector/matrix class library, according to user preference

0-521-75033-4, Hardback, \$70.00

Visit [us.cambridge.org/numericalrecipes](http://us.cambridge.org/numericalrecipes) for more information on the complete line of Numerical Recipes products.

Available in bookstores or from



**CAMBRIDGE**  
UNIVERSITY PRESS

800-872-7423

[us.cambridge.org/mathematics](http://us.cambridge.org/mathematics)

### Other new Numerical Recipes products for your library...

#### Numerical Recipes Example Book [C++]

0-521-75034-2, Paperback, \$35.00

#### Numerical Recipes in C and C++ Source Code CDROM with Windows, DOS, or Macintosh Single Screen License

0-521-75037-7, CD-ROM, \$50.00

#### Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras  
0-521-75036-9, CD-ROM, \$150.00

#### Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras  
0-521-75035-0, CD-ROM, \$90.00

# Numerical Recipes in Fortran from Cambridge University Press

## Numerical Recipes in Fortran 77

Volume 1 of Fortran Numerical Recipes  
Second Edition

*William H. Press, Saul A. Teukolsky,  
William T. Vetterling, and Brian P. Flannery*

"This reviewer knows of no other single source of  
so much material of this nature. Highly recommended."

—*Choice*

"...a valuable resource for those with a specific need for  
numerical software. The routines are prefaced with lucid, self-  
contained explanations...highly recommended for those who  
require the use and understanding of numerical software."

—*SIAM Review*

1992 992 pp. 0-521-43064-X Hardback \$70.00

### *Highlights include:*

- A chapter on integral equations and inverse methods
- Multigrid and other methods for solving partial differential equations
- Improved random number routines
- Wavelet transforms
- The statistical bootstrap method
- A chapter on "less-numerical" algorithms including compression coding and arbitrary precision arithmetic.

## Numerical Recipes in Fortran 77 Example Book

Second Edition

*William T. Vetterling, Saul A. Teukolsky, William H. Press, and Brian P. Flannery*

1992 256 pp. 0-521-43721-0 Paperback \$35.00

## Numerical Recipes in Fortran 90

The Art of Parallel Scientific Computing  
Volume 2 of Fortran Numerical Recipes  
Second Edition

*William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery*

"This present volume will contribute decisively to a significant breakthrough, as it provides models not only of the numerical algorithms for which previous editions are already famed, but also of an excellent Fortran 90 style."

—*From the Foreword by Michael Metcalf, one of Fortran 90's original designers and author of FORTRAN 90 Explained*

"This book is a classic and is essential reading for anyone concerned with the future of numerical calculation. It is beautifully produced, inexpensive for its content, and a must for any serious worker or student."

—*Computing Reviews*

Contains a detailed introduction to the Fortran 90 language and to the basic concepts of parallel programming, plus source code for all routines from the second edition of Numerical Recipes.

1996 576 pp. 0-521-57439-0 Hardback \$50.00

## Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras

2002 0-521-75036-9 CD-ROM \$150.00

## Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras

2002 0-521-75035-0 CD-ROM \$90.00

Visit [us.cambridge.org/numericalrecipes](http://us.cambridge.org/numericalrecipes) for more information on the complete line of *Numerical Recipes* products.

Available in bookstores or from



**CAMBRIDGE**  
UNIVERSITY PRESS

800-872-7423

[us.cambridge.org/mathematics](http://us.cambridge.org/mathematics)

# NEW! XML PLUG-IN FOR sas®

**IMPORT XML-FORMATTED DATA INTO SAS  
EXPORT SAS DATA AS XML-FORMATTED DATA**

Now, an easy way to move CDISC and other XML-formatted data into or out of your SAS-based systems. You don't have to know perl, XSLT, Xpath, Java®, or exotic languages. The remarkable Tekoa™ XML plug-in does it all for you.

Provided free of charge to any SAS user currently wrestling with XML. Developed by Zurich Biostatistics, the pioneer in SAS/XML integration.

## **FREE. EASY. AND IT WORKS.**

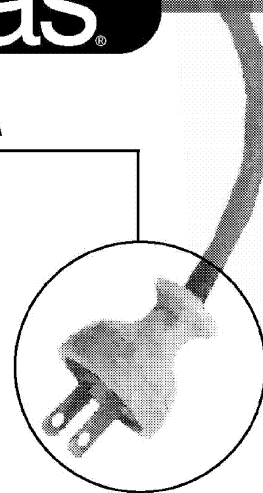
Just e-mail Michael Palmer (mcpalmer@zbi.net) and receive the fully-functional, proven Tekoa XML plug-in by e-mail.

*No charge. No obligation. No hassle. (We even support the tool. Imagine.)*

## **Zurich Biostatistics, Inc.**

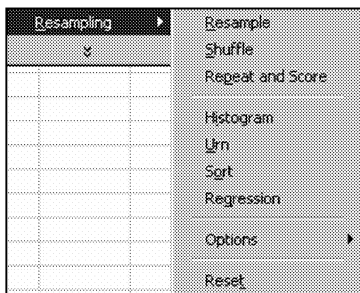
45 Park Place South, PMB 178, Morristown, NJ 07960 973 727-0025 www.zbi.net

***XML is easy if you know how. And we do.***



Tekoa XML Technology is a service mark of Zurich Biostatistics, Inc. SAS is a registered trademark of SAS Institute Inc. Java is a registered trademark of Sun Microsystems, Inc.

# Resampling Stats for Excel



Select the data you want to resample, select "resample" or "shuffle," then specify an output range for the resampled data. Calculate a statistic of interest, select "Repeat & Score," and the resampling operation will be repeated thousands (or tens of thousands) of times, and each time the value of your statistic of interest will be recorded. Does not use Excel's random number generator.

**View complete user guide and download free  
30-day trial at <[www.resample.com](http://www.resample.com)>**

\$249 commercial • \$149 personal/academic • \$89 student

612 N Jackson St., Arlington, VA 22201  
Tel 703-522-2713 • Fax 703-522-5846  
[stats@resample.com](mailto:stats@resample.com)



## ANNOUNCING LOGXACT 4

**Only LogXact 4 can fit a logistic regression model to the data on this page —even LogXact 2 cannot do it.**

To solve this problem, you need a very powerful exact logistic regression algorithm. LogXact 4 implements a ground-breaking network Monte Carlo algorithm (published in *JASA*, April 2000), extending the scope of LogXact to problems previously beyond its capacity.

PLUS! LogXact 4 also provides exact Poisson regression (used extensively for cohort studies in epidemiology).



### Take the Cytel Challenge

Data were gathered on 2,493 hospitalized patients, of whom 60 suffered from *clostridium difficile*, an acute form of diarrhea. Of interest was the relationship between the occurrence of diarrhea and age, length of hospital stage, sex, use of the antibiotic Clindomycin, and the use of the antibiotic Cephalexin.

When you have data like these (low response rates,

unbalanced covariates or small sample sizes), traditional logistic regression methods used by standard statistical packages often fail. But new LogXact 4 can fit models, test parameters and estimate coefficients even when the maximum likelihood method used by most statistical software can't. Plus, since LogXact gives you exact answers instead of approximations, it protects you from Type-I error.

#### 60 cases of diarrhea among 2,493 hospitalized patients

	Group 1	Group 2	•	Group 18
Cephalexin	0	0	•	1
Clindomycin	0	0	•	0
Sex	1	1	•	0
Age	0	0	•	1
LOS	0	1	•	1
<b>Diarrhea/Total ( 60/2,493)</b>	<b>0/174</b>	<b>1/113</b>	<b>•</b>	<b>4/4</b>

**For the full data set, the solution and a free 30 day trial of LogXact 4, visit [www.cytel.com](http://www.cytel.com)**

**Call (617) 661-2011; fax (617) 661-4405; e-mail: [sales@cytel.com](mailto:sales@cytel.com)**

CYTEL Software Corporation • 675 Massachusetts Ave., Cambridge, MA 02139 USA  
Tel (617) 661-2011 • Toll Free (US) 866-298-3511 • Fax (617) 661-4405  
<http://www.cytel.com> • E-mail: [sales@cytel.com](mailto:sales@cytel.com)

INTERNATIONAL DISTRIBUTORS: Ask Int'l (UK) e-mail: [cyteluk@asru.com](mailto:cyteluk@asru.com) • Tel: +44(0) 1227 795 240 • Fax: +44(0) 1227 795 201; ID2 (Belgium)  
• Tel: 32 2 6468918 • Fax: 32 2 4468662; Spadille Biostatistics ApS (Denmark) • [spadille@spadille.dk](mailto:spadille@spadille.dk) • Tel: 4548.484100 • Fax: 4248484200

**Cytel**  
STATISTICAL SOFTWARE



## JOIN DIVISION 5 OF APA!

The Division of Evaluation, Measurement, and Statistics of the American Psychological Association draws together individuals whose professional activities and/or interests include assessment, evaluation, measurement, and statistics. The disciplinary affiliation of division membership reaches well beyond psychology, includes both members and non-members of APA, and welcomes graduate students.

Benefits of membership include:

- subscription to *Psychological Methods* or *Psychological Assessment* (student members, who pay a reduced fee, do not automatically receive a journal, but may do so for an additional \$18)
- *The Score* – the division's quarterly newsletter
- Division's Listservs, which provide an opportunity for substantive discussions as well as the dissemination of important information (e.g., job openings, grant information, workshops)

Cost of membership: \$38 (**APA membership not required**); student membership is only \$8

For further information, please contact the Division's Membership Chair, Yossef Ben-Porath ([ybenpora@kent.edu](mailto:ybenpora@kent.edu)) or check out the Division's website:

<http://www.apa.org/divisions/div5/>

---

## ARE YOU INTERESTED IN AN ORGANIZATION DEVOTED TO EDUCATIONAL AND BEHAVIORAL STATISTICS?

Become a member of the **Special Interest Group - Educational Statisticians** of the American Educational Research Association (SIG-ES of AERA)!

The mission of SIG-ES is to increase the interaction among educational researchers interested in the theory, applications, and teaching of statistics in the social sciences.

Each Spring, as part of the overall AERA annual meeting, there are seven sessions sponsored by SIG-ES devoted to educational statistics and statistics education.

We also publish a twice-yearly electronic newsletter.

Past issues of the SIG-ES newsletter and other information regarding SIG-ES can be found at <http://orme.uark.edu/edstatsig.htm>

To join SIG-ES you must be a member of AERA. Dues are \$5.00 per year.

For more information, contact Joan Garfield, President of the SIG-ES, at [jbg@umn.edu](mailto:jbg@umn.edu).



# Lahey/Fujitsu Fortran

The standard for Fortran programming  
from the leader in Fortran language systems

**SOFTWARE SOLUTIONS**  
for Science & Engineering

## LF95 Fortran for Linux and Windows

Full Fortran 95/90/77 support  
Unsurpassed diagnostics  
Intel and AMD optimizations

IMSL compatible  
Fujitsu SSL2 math library  
Wisk graphics package

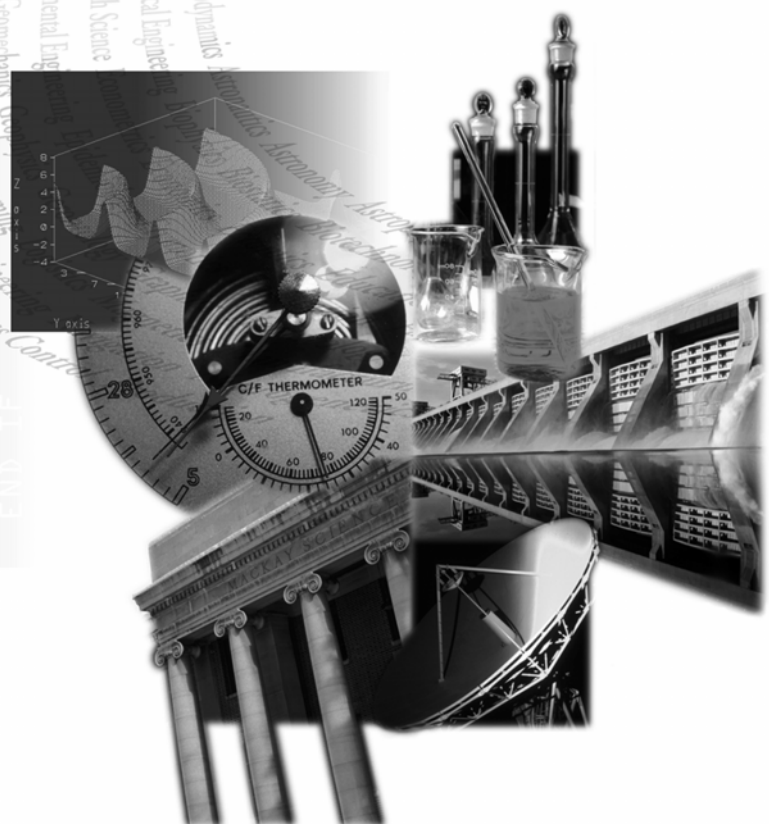
## LF Fortran for the Microsoft® .NET Framework - Coming Soon !

Visual Studio integration  
Windows / Web Forms designer  
Project and code templates

On-line integrated help  
XML Web services  
ADO.NET support

Visit [www.lahey.com](http://www.lahey.com) for more information

```
ELSE
  poly_coef
END IF
ELSE
  poly_coef
END IF
END FUNCTION poly_c
SUBROUTINE poly_ini
TYPE(poly), INTENT
REAL(fpkind), INTE
IF ( .NOT. PRESENT
  NULLIFY ( p%coef
ELSE
  m = UBOUND(v,i)
  IF ( max_degree
  ALLOCATE ( p%
  p%coeffs
ELSE
  ALLOC
  p%coeffs
END IF
END IF
```



Lahey Computer Systems, Inc.  
865 Tahoe Blvd - P.O. Box 6091  
Incline Village, NV 89450 USA  
1-775-831-2500  
[www.lahey.com](http://www.lahey.com)

## Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. The most recent American Psychological Association style guidelines are preferred.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at [ea@edstat.coe.wayne.edu](mailto:ea@edstat.coe.wayne.edu). Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are not acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font. If the technical expertise is available, submit the manuscript in two column format.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs “in-line”, not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified with indent.
8. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
9. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use “&” instead of “and” in multiple author listings.
10. *Suggestions for style*: Instead of “I drew a sample of 40” write “A sample of 40 was selected”. Use “although” instead of “while”, unless the meaning is “at the same time”. Use “because” instead of “since”, unless the meaning is “after”. Instead of “Smith (1990) notes” write “Smith (1990) noted”.

### Print Subscriptions

Print subscriptions including postage for professions is US \$60 per year; graduate students is US \$30 per year; and libraries, universities, and corporations is US \$195 per year. Subscribers outside of the US and Canada pay a US \$10 surcharge for additional postage. Online access is currently free at <http://tbf.coe.wayne.edu/jmasm>. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to [jmasm@edstat.coe.wayne.edu](mailto:jmasm@edstat.coe.wayne.edu).

### Notice To Advertisers

Send requests for advertising information to [jmasm@edstat.coe.wayne.edu](mailto:jmasm@edstat.coe.wayne.edu).